THE UNIVERSITY
of EDINBURGH

LitCrit: Exploring intentions as a basis for automated feedback on Related Work

Arlene J Casey

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
The University of Edinburgh
2020

Abstract

Learning the skill of academic writing is critical for post-graduate (PG) students to be successful, yet many struggle to master the required standard. Feedback can play a formative role in developing these skills, but many students do not find sufficiently helpful the kinds of feedback available to them. As the *Related Work* section is known to be particularly difficult for PG students to master that is the focus of this thesis.

To date, models of academic writing have been built on observational studies of academic articles. In contrast, we carry out a user study to explore what content experts look for in *Related Work* and how this differs from PG students. We claim that by understanding what experts look for in *Related Work* and what aspects PG students struggle with, a useful author intention model can be developed to support writing feedback for *Related Work* sections. Our work demonstrates reliable annotation of the model intentions. Developing on existing algorithms, designed to identify rhetorical intentions in academic writing, we build a supervised machine learning classifier, showing how features focused on *Related Work* sections improve recognition of content aspects. Carrying out a study to rate the quality of *Related Work*, we demonstrate that the model is a good proxy for predicting quality, validating the choice of intentions in our model. In addition to recognising author intentions, we automate the generation of feedback based on observations of intentions that are present and missing, taking into account areas that PG students struggle to recognise.

The thesis also contributes a new prototype writing analytic tool, called **LitCrit**, that supports visualising the intention narrative of *Related Work* and presents feedback. We claim this visualisation approach changes the PG student's perception of *Related Work*, and demonstrate through a user study that it does draw attention to aspects previously missed bringing PG student responses in line with experts. Finally, we explore the performance of our classifier, originally set within the Computational Linguistics discipline, to that of Computer Graphics. This shows us that while performance may be lower when care is taken to understand those features which are discipline dependent, there is scope for improvement. Also, while a discipline may have the same intentions present in a section, their structural presentation may differ impacting feature choice.

Lay Summary

To be successful students must learn the skill of academic writing. It is needed to complete assignments, submit a thesis and publish papers. Many students struggle to master the required standard. Students develop this skill through feedback from supervisors and peers. However, often what students receive is not adequate as it focuses on aspects of grammar and spelling with no pointers about non-existent or problematic content. Recently, online systems that help students with their writing have become popular, particularly those that go beyond the basics of grammar and spelling, focusing on aspects of content that are expected but may be missing. One area that has not been addressed by these tools, but in which students are known to struggle, is in writing a *Related Work* section – where an author writes about previous research and its relationship to their own work. Our motivation in this thesis is the idea of supporting students by providing automated feedback on their *Related Work* writing.

To help students with their writing, we must understand what content should be present within the *Related Work* section and provide feedback on whether this content is present or missing. Our work differs from previous work that seeks to understand content in that we use a peer-review exercise, rather than observational studies of existing published articles. Peer-review is known to be challenging as experts often disagree on content aspects. Nonetheless, we can show that experts do agree on what content should be present within *Related Work*. In addition, we compare expert peer-review to peer-review by students. This comparison allows us to understand how students differ and what aspects of writing they struggle with most. Using what we learn about content and where students struggle, we build a model to support writing feedback presenting this within **LitCrit**, a tool developed as part of this thesis. This tool highlights the *Related Work* narrative, drawing attention to the presence or absence of content and providing feedback. Through user studies, we show that using **LitCrit** draws students' attention to aspects of content they previously overlooked, helping them think more critically about the writing. We show that there is a relationship between our model of author intentions and quality scores given to *Related Works*. This relationship validates our model showing that it does indeed represent expected content within *Related Work*. Finally, we explore the applicability of our model to another scientific discipline, Computer Graphics, showing how the difference between the disciplines impacts the performance of being able to recognise content.

Acknowledgements

Many people have helped in extraordinary ways to support me during this journey, but the person who has profoundly impacted my PhD journey is Amy Isard. What were the chances two Scottish, mature PhD students, who like cats would find each other at a machine learning school in Portugal? That chance encounter introduced me to Informatics@Edinburgh, where I was initially a visiting student, and Amy encouraged me in taking the plunge and transferring Universities mid-way. Amy has shared her knowledge, her friends and helped me integrate at Edinburgh, but also been there when I needed to talk about my research and first to volunteer when I needed an experiment done or a paper read. Thank-you!

I am sincerely grateful to all my supervisors, of which there are a few. Samad Ahmadi and Fionn Murtagh at De Montfort University. Samad's encouragement and belief in me to undertake a PhD. After leaving De Montfort, Fionn went out of his way to provide me with guidance and always made the time to support my research and my transition to Edinburgh. On transferring to Edinburgh, I remember thinking I had won the lottery when both Jon Oberlander and Bonnie Webber said they were interested in co-supervising me. Jon's enthusiasm and interest were unparalleled and made every meeting something to look forward to. His untimely passing was a great shock and loss to all who knew him. Bonnie's wisdom and experience in giving me the freedom to explore and find my own path in this research, but also her judgement in knowing when to guide me in the right direction has been greatly appreciated. Agreeing to join a co-supervision team mid-way through a PhD is not an easy supervision choice, and I appreciate that Dorota Głowacka embraced this so readily and stayed with me when she left for Helsinki. Both Bonnie and Dorota have given their time freely, sharing their knowledge and providing me with insightful discussion and encouragement, which has helped develop and shape me as a researcher. I will always be thankful to them for this opportunity to learn from them.

Finally, my friends and family who have always supported and encouraged me. My new friends at Edinburgh who were willing to talk about my research and volunteer for experiments. My long-standing friends who have been understanding and forgiving when I spent too much time on my thesis and not enough time with them. My parents and siblings with their unconditional belief in me. And Gerry, your patience and understanding has been tested as I took on 'my mid-life crisis project', but your support and faith in me has been unwavering. It is truly appreciated.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Arlene J Casey)*

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Academic researchers seek to find, understand and critically review others' research through published scientific articles. This not only helps them to learn, be motivated and inspired but allows them to position their work within the field, justifying its need and qualifying their papers for publication. Academic writing is, therefore, a critical skill for post-graduate (PG) students to learn as it is used in assessment by peers and to publish work. However, writing is an aspect that PG students often find difficult (Aitchison et al., 2012; Ross et al., 2011) particularly in areas, such as Computer Science, where their interest is more in developing technical skills rather than writing. Beyond the basics of writing, spelling and grammar, PG students must grasp expectations of language, style and content structure in academic writing. Evidence shows though that PG students struggle to identify and learn the practices that are expected in quality writing (Aitchison et al., 2012; Paltridge and Starfield, 2007).

While the motivation behind this work is to provide PG students with support in academic writing, good writing has multiple aspects. The focus of this thesis is on one aspect of writing, rhetorical intentions, also known as author intentions. These are conventions of structures and arguments expected to be present within the writing. Existing work that proposes models to represent academic text and support writing are built on observations of published articles. We claim that using peer-review from experts about content that is present and missing, rather than observational studies, can allow for the development of an author intention model to support writing feedback. In addition to using peer-review from experts, we also use that of PG students, learning what aspects they particularly struggle with and incorporating this into the model. If PG students struggle to recognise aspects in others' writing, it is likely they will miss this in their own writing. We validate our author intention model by showing it is a

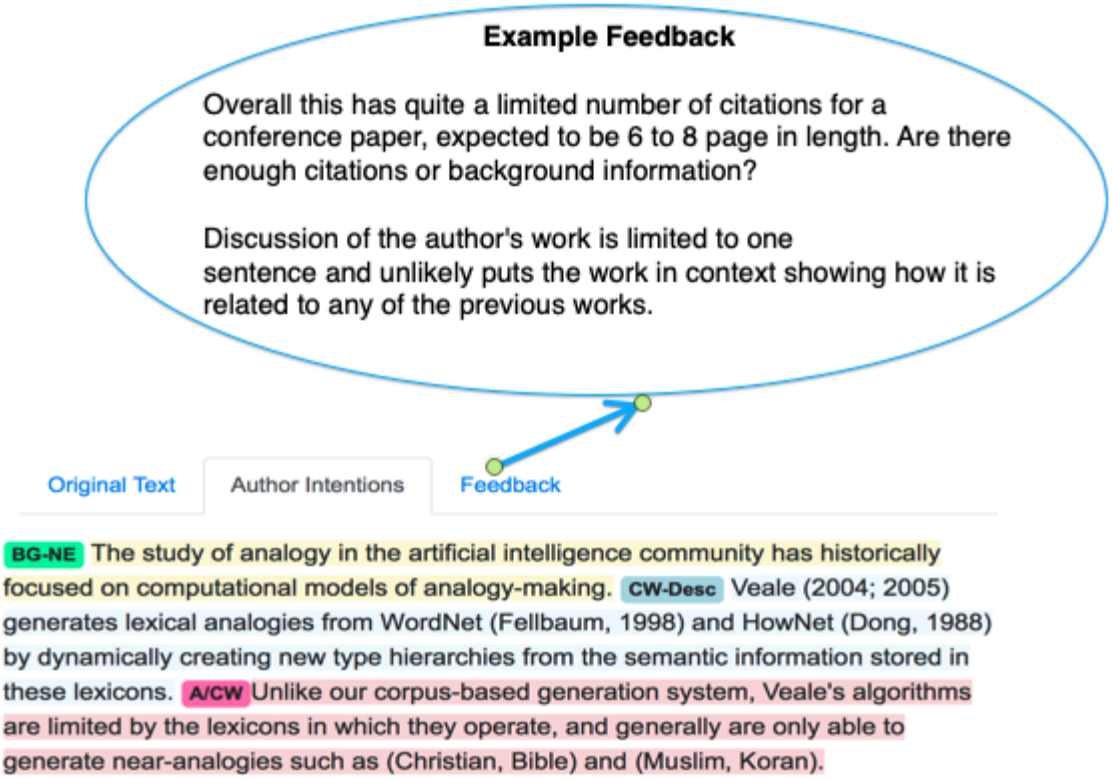


Figure 1.1: Example of a *Related Work* with feedback that highlights what may be missing

good representation for content by using it to predict the quality of the writing. Using our author intention model to highlight the narrative of writing we claim will draw a PG student's attention to aspects of content they previously missed changing their perceptions and thinking about the writing, bringing them more in line with experts. We find evidence for this through our user study with PG students using **LitCrit**, an analytic writing tool developed as part of this thesis, which highlights the narrative with our author intention model. Figure 1.1 gives an example of *Related Work* writing and highlighting of the rhetoric structure, with feedback, that makes suggestions that could help improve the writing. In this example we see that there is not enough related work discussed and the feedback suggests they may not have put the related work in context to their own work.

Due to the challenges in working with scientific writing, such as linguistic variation across disciplines, the work here is limited to the domain of Computational Linguistics. However, Chapter 7 does assess how transferable the framework and model developed is to another domain, using examples from the discipline of Computer Graphics. In addition, only rhetoric intentions within a *Related Work* section – what an author says

about others' work and its relation to their own work – are considered. The reasons for this are described in the next section.

1.1 Why a *Related Work* Section?

Related Work is an area that PG students are known to struggle with, not just because they must learn the structures and intentions that should be present but they must find and project their viewpoint a challenge for PG students when learning to write (Kamler and Thomson, 2006). Often, lack of experience and confidence in projecting their voice amongst established scholars results in students making bland statements about others' work. This is a view supported by Boote and Beile (2016) who discuss the decline of the literature review section in PhD theses. Such bland statements in the *Related Work* section do nothing more than provide a list of work with no real critical commentary or attempt to relate it to the author's own work. Cited works should be ones that have implications for the author's work (Maxwell, 2006). Swales (Swales, 1981, 1990) discusses the difficulties in writing for introductory material including previous work and that this is much harder than *Results* and *Methods* sections. As a result of being harder to write there is less focus on teaching these skills to PG students, and supervisors themselves find it hard to give adequate feedback on how to improve such sections.

Related Work sections do not occur in every discipline. Some disciplines tend to have their literature work in the *Introduction* section, e.g. some life-science sub-fields. However, in the domain chosen for this thesis, *Related Work* sections are much more prevalent. Whether a PG student works within a discipline that has a *Related Work* section in an article or not, there is still a need to understand how to write about others' work and put this in perspective to their own work. Almost all thesis work requires literature work and even if an article does not contain a specific *Related Work* section, there is an expectation that there will be some mention of how the author's work relates to others' work. This makes understanding how to form intentions in writing about others' work and how this relates to their own a valuable skill for PG students to learn, regardless of whether their discipline has specific *Related Work* sections or not. Several tools exist to support academic writing but none of these specifically focus on a *Related Work* section.

1.2 Using Author Intentions for Automated PG Writing Support

In this thesis we are motivated by the idea of helping PG students with their academic writing but our focus is on one aspect, rhetorical intentions expected to be present in writing. This is challenging, though, as academic writing has multiple aspects that combine to produce expected content. Some aspects are more objective and can be measured, such as grammar or correct citation use. Others are more subjective, such as how to creatively use language to meet readers' expectations about what structures and intentions should occur, and this often requires discipline knowledge. Studies have shown that language used can differ across the typical sections found in a research article (*Introduction, Methods, Results, Discussion*) and different disciplines have accepted conventions within their writing, meaning linguistic constructs and section content can differ across these disciplines (Hyland, 2015). PG students must learn the socially accepted norms and language of their discipline but the subjectivity of such measures make them much harder to learn and automate as they are not exact (O'Rourke and Calvo, 2009), requiring in-depth analysis with advanced text-mining techniques.

By far, the most significant development of PG writing skills comes through feedback from supervisors or peer-review (Abel et al., 2018). However, feedback often merely involves notes scribbled on printouts, and even feedback from experienced supervisors may be unclear and unhelpful (Paré, 2010; Aitchison et al., 2012). Students feel frustrated with the feedback given by supervisors. Students can be left waiting for months, and some supervisors only focus on aspects of grammar and structure, rewriting with no explanation and no provision on how to write (Aitchison et al., 2012). Work by Ross et al. (2011) demonstrates this frustration with one student stating *"feedback was not helpful, mainly about grammar, not about how to write in an expanded form"*. Supervisors often complain though that PG students lack familiarity with the patterns of writing expected within their discipline or lack an awareness of the audience they are writing for (Maher et al., 2014). This insight into the disciplinary discourse community is necessary, and PG students need to develop an understanding of the rhetorical aspects of academic writing (Abel et al., 2018) and its overall purpose, presenting the researchers thinking. There is pedagogical value to be found in helping PG students become aware of intentions and structures that should be used, enabling them to identify and learn patterns expected in their discipline. Swales (1990) argues that a

student's understanding of these rhetoric structures is likely to be as important as an understanding of grammar.

Recent years have seen more focus on building tools, known as writing analytic tools, to include subjective measures, such as rhetoric intentions to provide more formative and actionable feedback on a student's writing. These tools use visualisation to highlight aspects of intentions in writing, drawing the attention of the writer and helping them to see how their thinking is represented in their writing, or where aspects may be missing. In addition, these types of tools provide instant feedback with no need to wait for a supervisor's response. However, the work that focuses on rhetoric intentions for PG level writing mainly address *Introductions* or *Abstracts* (Cotos and Pendar, 2016; Anthony and V. Lashkia, 2003; Feltrim et al., 2006; Abel et al., 2018). Most likely this focus on *Introductions* is due to the formative work of Swales (1981, 1990) who developed the Create a Research Space (CARS) model that is based on observational studies of rhetoric intentions within an *Introduction* section to support writing. The focus on sections other than *Related Work* means that these intention models miss aspects that are relevant for writing support within a *Related Work* section or have intentions that are irrelevant.

1.3 Overview of Thesis Research

This thesis is based on two main claims:

- Claim 1: Peer-review from experts, and identifying where PG students struggle through peer-review, can be used to model author intentions that represent what content should be present in a *Related Work*.

- Claim 2: Highlighting the narrative with our author intention model can influence PG students' thinking and perception of *Related Work*, bringing their views in line with experts.

This section gives an overview of the work undertaken and the motivation behind this to provide evidence to our claims. The research questions listed here are revisited in Chapter 2 with a more in-depth discussion and how they contribute to a gap in current knowledge. We end with a summary of our main contributions.

1.3.1 Content that Should be Present in *Related Work* Sections

The following research questions are investigated in this part of the work:

1. What are the content expectations highlighted in a *Related Work* by experts and do experts agree with each other?

2. Do PG students differ from experts in what they look for in a *Related Work*?

There are considerable resources that could be used to define expected content in academic writing, such as *Helping doctoral students write: Pedagogies for supervision* (Kamler and Thomson, 2006), and *The Craft of Communication* (Harmon and Gross, 2010). However, there is no evidence that these are the same characteristics that experts actually use when conducting peer-review. Disagreement at the peer-review stage of published papers is a widely known issue. For example, a controlled experiment involving peer-review of papers submitted to NIPS 2014 - a leading machine learning conference - showed that two committees disagreed on accept/reject decisions in over 25% of papers (Lawrence and Cortes, 2014).

In contrast to previous approaches of academic writing tools, which use intention models based on observational study of academic writing, this thesis uses an intention model based on a study of peer-review given by experts. We undertake a user study specifically looking for where there is agreement by experts in the intentions highlighted. Additionally, peer-review given by experts is compared to that of PG students. Novices undertaking peer-review have been shown to have difficulties differentiating good work from bad and often lack the in-depth perspective required to identify what aspects matter (Cambre et al., 2018). It is likely that if PG students cannot see this in others' work, they may miss it when describing their own work. The intentions proposed in this thesis are designed to focus strongly on aspects PG students are found less likely to identify compared to experts.

1.3.2 Building a Model of Author Intentions for Writing Support

The following research questions are investigated in this part of the work:

1. Can the author intention labels be annotated with reasonable human agreement?

2. Can the author intention labels be recognised automatically with reasonable accuracy?

3. Do the author intention labels serve as a proxy for indicating content quality in a *Related Work*?

A model of author intention labels is proposed based on the results from the outcome of our peer-review study. Successful automated recognition of author intentions have been based on robust annotation studies but ambiguity is known to be a problem in human agreement on annotation in scientific publications (Stab and Gurevych, 2014; Kirschner et al., 2015). Carrying out an annotation study with our author intention model good agreement between humans is found for the author intentions with inter-annotator agreement of 77%. Previous approaches have shown reasonable success using machine learning approaches to automate the recognition of author intentions. However, to give feedback in writing, recognition rates need to be both high and consistent, reaching that of human agreement. The annotated data is used to train a supervised machine learning model to recognise author intentions, which in its final iteration reaches 76.34%, almost that of human agreement.

The focus of this work is to help PG students realise intentions that are present or missing in their writing and a prediction of quality is not particularly helpful when writing. The model of author intention is designed to represent what content experts expect to be present. Quality prediction may not be helpful for writing but it is useful to explore how well the intentions chosen represent the overall quality of *Related Work*. An assessment study to rate the quality of the *Related Works* is first undertaken, and then the author intention labels are used to show good prediction of quality for a *Related Work*. Also explored is the different occurrence of the author labels within the quality rated *Related Works*. This work supports the earlier findings that the author intentions are representative of quality and thus expected content, with poorer *Related Works* showing significant differences in intention label occurrence than good *Related Works*.

1.3.3 Evaluating the Visualisation of Author Intentions

The following research questions are investigated as part of this work:

1. Does highlighting the narrative structure with intentions change PG student perceptions of a *Related Work*?

2. Do PG students find the visualisation of intentions and feedback on missing aspects helpful?

A user study with PG students is carried out this time highlighting intentions and providing feedback on the writing. To perform this study we use **LitCrit**, a writing analytic tool developed as part of this thesis. In addition to highlighting intentions, **LitCrit** provides feedback on aspects that are present and may be missing. To do this, we develop a discourse segmentation method for *Related Work* that enables us to identify contextual feedback to give. We show that the performance of our segmentation is high compared with human segmentation. The PG student responses after using **LitCrit** are compared to those in the previous peer-review study without **LitCrit**. Findings show that using **LitCrit** changes the PG students opinion of the *Related Work* bringing these in line with experts and the evaluation of **LitCrit** is overall positive.

1.3.4 Investigating Discipline Independence

The final aspect of our work considers if the recognition of author intentions within a *Related Work* can be applied within another discipline, the discipline of Computer Graphics. Previously mentioned are the problems of linguistic variation between disciplines and the likelihood that methods for recognising features in one discipline will not work in another. Findings show that automated recognition of intentions can be achieved although it is significantly lower and requires dropping some of the more domain specific features.

1.4 Research Contributions

The main contributions of this thesis can be broadly grouped into two themes. The first area of contribution this thesis centres around is understanding what content should be present in a *Related Work* and how this can be used to build a model of author intention that represents expected content. The contributions in this part would be of interest to the Education research community. The contributions can be summarised as follows:

- This research is the first study to use peer-review of experts to develop an author intention model. We demonstrate through this study that experts agree on content that should be present in a *Related Work*. In addition, we identify how PG students differ from experts in recognising arguments that should be present within *Related Work*.

- Based on the content that experts expect to see, and taking into account areas PG students struggle with, we develop a model of author intentions that can be used to represent expected content in a *Related Work* that supports writing feedback.

- We validate our model of content experts expect to see when we show that the author intention labels are good predictors of the quality of a *Related Work* with poor *Related Works* missing vital author intentions.

- We develop a prototype tool **LitCrit** and using this we show that highlighting the visual narrative of intentions, within *Related Work* writing, influences a PG student's perception of a *Related Work*, bringing this more in line with an expert.

The second contribution area of the thesis focuses on the NLP components that support the automatic recognition of our author intention model and demonstrate its viability. The contributions here would be of most interest to the NLP research community. The thesis contribution can be summarised as follows:

- Through an annotation study, we show that the model of author intentions can be reliably annotated by humans.

- We build a model based on supervised machine learning that can predict, with almost human accuracy, our author intention labels. We show that directing attention to local features found in one section and using between sentence context can improve the performance of the classifier.

- To provide writing feedback, we propose and evaluate a method that automates segmenting the text within a *Related Work*. This allows for context about the cited work and author's work to be understood, providing feedback comments in addition to visualising the intention narrative.

1.5 Thesis Outline

This thesis is organised as follows:

Chapter 2 This chapter presents a background on previous work in recognising author intentions within scientific publications and describes how these influence the approach taken in the thesis. Also discussed are other writing analytic tools explicitly aimed at academic writing. Aspects of writing other than intentions,

e.g. structure, grammar, are raised and how these are important in writing feedback but not within the scope of this thesis. What makes a good *Related Work* is discussed as this helps to form the design of the study in Chapter 3. Explored further is why PG students struggle with writing about work related to their own. The chapter concludes with a summary of current work and puts our work in context to the gaps in current research along with our hypotheses and our research questions.

Chapter 3 Investigates through a peer-review study what content should be present in *Related Work*. This chapter describes the user study presenting the aspects of content experts deem necessary within a *Related Work* and highlighting particular aspects PG students struggle with.

Chapter 4 Builds a model of author intention to represent content aspects that should be present in *Related Work*. We discuss how this relates to other intention models that have been proposed within academic writing and present the results from our annotation study. Results from this chapter have been published in (Casey et al., 2019b) in the 13th Linguistics Annotation Workshop.

Chapter 5 Builds a supervised classifier to automatically recognise the author intention labels in *Related Work* writing. Error analysis is provided to give insight into problems the classifier encounters and how such problems may provide insight into supporting writing feedback. Further experiments are described based on the error analysis that improve the classifier performance. Results from this chapter have been published in (Casey et al., 2019c) in the RANLP Conference.

Chapter 6 Investigates through a user study how the highlighting of author intentions changes a PG student's perception of a *Related Work*. The chapter describes the design and functionality of the writing analytic tool **LitCrit**, developed as part of this thesis. We describe our automatic discourse segmentation to consider context and generate feedback on the missing aspects, and the accuracy of this is evaluated.

Chapter 7 Investigates the discipline independence of the author intention model, applying it to the discipline of Computer Graphics. We discuss the modifications to the feature set that enable better performance of the classifier in this discipline.

Chapter 8 Investigates how well the author intention model represents the overall

quality of the *Related Work* section. We describe the assessment carried out to rate the *Related Work* sections and the experiments undertaken to use the author intentions to predict quality. Results from this chapter have been published in (Casey et al., 2019a) in the BIRNDL Workshop.

Chapter 9 Provides a summary of the work undertaken, the results and discusses some of the limitations and avenues for future work that can build on the results of this thesis. Also highlighted are some of the outcomes that could prove useful from a pedagogical perspective.

Chapter 2

Background

2.1 Introduction

This chapter gives background and discusses other works that motivate and have been used in this thesis. There are several areas addressed in this chapter.

First, we consider other models of author intentions. This work seeks to model the intentional structure of a *Related Work*. Whilst no other model exists that addresses this directly there are aspects of other models that are relevant, e.g. those that address citation function, some of the steps described in Swales CARS model (Swales, 1990) or some zones from Argument Zoning (Teufel, 1999) or relevant work within Argument Mining.

Secondly, we discuss the field of automated writing evaluation and current state-of-the-art methods in this field. We also describe specifically some current writing analytic tools based on intention which are specifically aimed at academic writing are described - AcaWriter[1], Research Writing Tutor[2], Criterion[3].

Thirdly, although not within the scope of this thesis, other aspects of writing, such as structure, readability, grammar, which may have a bearing on the ability to automate the recognition of the intention labels are discussed.

Additionally, included is a section which discusses the *Related Work* section offering some insight into why this may be declining and the notion of what

[1] https://www.uts.edu.au/research-and-teaching/teaching-and-research-integration/acawriter
[2] https://cce.grad-college.iastate.edu/resources/writing-resources
[3] https://www.ets.org/criterion

makes a good *Related Work* which laid the foundation for the study in the next chapter.

Finally, we discuss the findings from each of these section and how they relate to the research questions and motivate the approach taken in this thesis.

2.2 Intention and Argument Modelling in Academic Writing

There is a large body of research on building frameworks to recognise intentions in academic writing but each of these is developed for specific purposes that do not necessarily align with writing support. Whilst these may not align with writing support directly, an understanding may prove valuable in developing the *Related Work* intention framework and best practices for automatically recognising intentions.

Swales' model is introduced first to help give a deeper understanding of what author intentions are. Direct comparison of intention labels of the most related models to the ones developed in this thesis is carried out in Chapter 3. Here, the models that are most closely related or motivated work in this thesis are described more generally, along with the current performance levels and methods being used in these models.

2.2.1 Author Intention Models

This section describes author intention models used within academic writing.

2.2.1.1 CARS Model

One of the earliest and most influential studies that attempted to model academic discourse structure is that of Swales (1981, 1990). Swales considered *Introduction* sections in research articles proposing that rhetoric intentions are specific to sections within an article. Intentions are realised with a sequence of moves or steps. Each of these steps is key in allowing an author to convey and provide a persuasive message, which forms the author intention. Swales carried

out an observational study on *Introduction* sections suggesting three main author intentions (Moves) exist – defining these in his Create a Research Space model (CARS). The model presented is from (Swales, 1990, p. 141). The previous model in (Swales, 1981), originally containing four moves, was modified to create this new model. Swales puts this modification down to revisions that took place as the original corpus of *Introductions* studied were shorter and the new proposed model allowed for patterns that occurred in longer *Introductions*. The CARS model is presented in Table 2.1 and consists of 3 main moves (intentions) **1 -Establishing a Territory**, **2- Establishing a Niche** and **3 - Occupying a Niche**, each with a number of steps. Reviewing the Moves and Steps in the CARS model we can see some would occur in a *Related Work*, such as in Move 2, Step 1B *Indicating a Gap* or Step 1D *Continuing a Tradition* or Move 1, Step 3 *Reviewing items of previous research*. However, the expert opinions found in Section 3.7 show that the CARS model does not cover all the intentions that are important in *Related Work*, e.g. putting the cited work in context as opposed to just listing cited works.

Swales shows how moves are linked to cue phrases and words, e.g. the move establishing a territory is often found with phrases, such as *it is well known that* or time-based phrases *recent studies...* and reporting verbs, such as *show, establish*. However, he argues that viable correlations between rhetoric and linguistic features can only be established within a genre where the language is sufficiently narrow and focused on a communicative purpose. Additionally, he points out that this narrowing of focus limits the ability to say anything useful outside of the genre or specific section of focus, e.g. narrowing in on an *Introduction* section limits what can be said about the *Results* section. This means that aspects of existing intention models may apply to a *Related Work* but others may be irrelevant or missing. Swales' model or aspects of it have been operationalised within existing writing analytic tools such as AcaWriter (Abel et al., 2018) and Research Writing Tutor (Cotos and Pendar, 2016), described in Section 2.3.

2.2.1.2 Argument Zoning Intentions

Perhaps one of the best used models of labelling rhetoric structure is Argument Zoning (AZ) (Teufel, 1999). This model labels sentences with argument zones representing the rhetoric intention of the sentence within the global context of

Swales CARS Model
Move 1 - Creating a territory
Authors describe motivation for a problem.
Step 1 - Claiming centrality
Step 2 - Making topic generalization(s)
Step 3 -Reviewing items of previous research
Move 2 - Establishing a niche
Puts forth the goal of the current research by identifying a gap in prior work or raising a question that needs to be solved.
Step 1A - Counter-claiming
Step 1B - Indicating a gap
Step 1C - Question-Raising
Step 1D - Continuing a tradition
Move 3 - Occupying the niche
Involves description of the new work and associated details.
Step 1A - Outlining purposes
Step 1B - Announcing present research
Step 2 - Announcing principal findings
Step 3 - Indicating research article structure

Table 2.1: Create a Research(CARS) Intention Model,(Swales, 1990)

AZ Class	Description
BKG	General Scientific Background
OTH	Neutral description of other people's work
OWN	Neutral description of the own, new work
AIM	Statements of the particular aim of the current paper
TXT	Statements of textual organization of the current paper (in chapter 1, we introduce ..)
CTR	Contrastive or comparative statements about other work; explicit mentions of weakness in other work
BAS	Statements that own work is based on other work

Figure 2.1: The seven Argument Zones from the AZ schema with the description of each zone in the right column (Teufel, 1999).

the document, e.g. background, aim or conclusion. The schema is designed to support document summarising and information extraction within the Computational Linguistic field. AZ is based on seven author intention labels, described in Figure 2.1. Three labels provide intellectual ownership about what is being discussed in a sentence : *Background, Other, Own*. Four labels provide more indication of the rhetorical move: *Aim, Basis, Contrast, Textual*. Each sentence is assigned a single zone label, and the author points out that large chunks of text are sometimes labelled as *Background*. This leads her to question the nature of this text, as space in a research article is of a premium why not say something more substantial, such as evaluate the work that is being cited. This ties in with earlier points about decline in literature work and cited work should be cited for a reason not just to provide a bland description.

Teufel highlights the challenges in annotating data in academic articles. Often the context is linguistically unmarked, which can make judgements about the relationship of the cited work more difficult. In Section 1.1 it was highlighted that novice writers could struggle to provide citations that go beyond lists or brief description, and this leads to what Teufel calls *linguistically unmarked context*. This subjective nature of determining the relationship of the cited work to that of the author(s) makes it hard to operationalise (Teufel, 1999; Swales, 1990). The reader's experience also has a role to play in the interpretation of function, with experts in the field not requiring as many linguistic clues to relevance as a novice

reader may require.

Teufel and Moens (2002) report an overall accuracy of 73% in automatic recognition of AZ labels, employing a Naive Bayes classifier. However, the Macro-F score is low at 50% with individual F-measures low for sparse categories, e.g. CONTRAST 26% and BASIS 38%. One of the problems of using Naive Bayes is that it requires conditional independence and the features used are not independent. In later work (Teufel and Kan, 2009), a Maximum Entropy model is used to predict the Argument Zones resulting in a lower accuracy of 66.80% and individual F1 scores still low for sparse labels: Own 81%, Aim 51%, Basis 22%, Background 24%, Contrast, 19%, Other 31%, Textual 61%. It is worth noting that in her original thesis work on AZ, Teufel uses aspects of the CARS model in developing her Argument Zoning. However, she rejects its direct use, calling it 'too hard to operationalise'.

The AZ schema was extended from 7 to 15 finer-grained categories, seen in Table 2.2, and called AZ-II (Teufel et al., 2009). Having extended the schema, they were able to apply it within another domain, Chemistry. The changes to the schema enable some of the differences in this field to be captured. For example, the new fine-grained OWN categories are needed for Chemistry but these are not as readily observed in the original discipline of Computational Linguistics. In addition, the original purpose of AZ was for summarisation and had no need for these subdivided categories. The AZ scheme has been successfully applied in other domains, e.g. biology papers (Mizuta and Collier, 2004) and within the astronomy discipline (Merity et al., 2009). However, like when applying AZ to Chemistry, the schema itself was modified to adapt to these domains. This highlights the differences that exist in content between disciplines and how this makes it difficult to create generic models that cover all disciplines.

2.2.1.3 Scientific Core Concepts Model (CoreSC)

Liakata et al. (2012) took a different approach to label author intentions, studying the conceptual structure of life science articles treating the article as a scientific investigation. This approach reflects the nature and layout of articles within the life-science domain. The schema is ontology motivated and creates 11 categories, as described in Table 2.3. Comparing support vector machine (SVM) and conditional random fields (CRF), they show the highest accuracy for the classi-

Category	Description
AIM	Statement of specific research goal, or hypothesis of current paper
NOV_ADV	Novelty or advantage of own approach
CO_GRO	No knowledge claim is raised (or knowledge claim not significant for the paper)
OTHR	Knowledge claim(significant for paper) held by someone else. Neutral description
PREV_OWN	Knowledge claim (significant) held by authors in previous paper. Neutral description
OWN_MTHD	New knowledge claim, own work: methods
OWN_FAIL	A solution/method/experiment in the paper that did not work
OWN_RES	Measurable objective outcome of own work
OWN_CONC	Findings, Conclusions (non measurable) of own work
CoDI	Comparison, contrast, difference to other solution (neutral)
GAP_WEAK	Lack of solution in field, problem with other solutions
ANTISUPP	Clash with somebody else's results or theory; superiority of own work
SUPPORT	Other work supports current work or is supported by current work
USE	Other work is used in own work
FUT	Statements/suggestions about future work (own or general)

Table 2.2: The AZ-II schema giving the 15 intention label categories and a description of each category (Teufel et al., 2009).

CoreSC Class	Description
Hypothesis	A statement not yet confirmed rather than a factual statement
Motivation	The reasons behind an investigation
Background	Generally accepted background knowledge and previous work
Goal	A target state of the investigation where intended discoveries are made
Object-New	An entity which is a product or main theme of the investigation
Method-New	Means by which an author seeks to achieve a goal of the investigation
Method-Old	A method mentioned pertaining to previous work
Experiment	An experiment method
Model	A statement about theoretical model or framework
Observation	The data/phenomena recorded in an investigation
Result	Factual statements about the outputs of an investigation
Conclusion	Statements inferred from observations & results relating to research hypothesis

Table 2.3: The 11 categories of intention labels in CoreSC, including the two sub categories for Method, and provides a description of each label(Liakata et al., 2012) .

fier is with SVM at 51.60%. Individual F1 scores: Hypothesis 19%, Motivation 10%, Background 62%, Goal 26%, Object 24%, Method 29%, Experiment 75%, Model 53%, Observation 50%, Result 51%, Conclusion 45%. CoreSC has more categories than AZ and thus more frequent lower scoring categories likely resulting in the overall lower score. However, the features used in classification are quite different from AZ. The features mainly focus on items like sentence location, counts, verbs, verb tense and n-grams. AZ features, discussed more in Section 5.3.1, make use of an extensive study of patterns that occur and relate these to lexicon categories. The AZ approach likely finds a better way to generalise the variation of language and structure that occurs.

ArguminSci Multi-Layer Labels

Annotation Schema	Tags
Scientific Discourse	Approach, Background, Challenge, Outcome, Future Work
Subjective Statements	Novelty, Common_practice, Advantage, Disadvantage, Limitation
Summary	Relevance for summary, this is a grade with respect to the relevance of this sentence for inclusion in a summary of the document
Citation Purpose	Criticism, Neutral, Comparison, Use, Basis, Substantiation

Table 2.4: The four schemas of Annotation for the ArguminSci models and a list of label tags in the right column for each layer. (Fisas et al., 2015)

2.2.1.4 ArguminSci Model

Fisas et al. (2015) developed a schema based on both AZ and CoreSC models to represent scientific concepts that appear in Computer Graphics articles. Their schema focuses on providing a standardised structure to track the progress in a scientific field. They propose four types of schemas each containing their own labels, described in Table 2.4. The labels are applied at a sentence level except for citation purpose, which captures a citation purpose across sentences. Fisas et al. (2015) show good recognition of intention labels with an Average F1 score of 80.10% using Logistic Regression on the Scientific Discourse layer labels. F1 scores: Approach 87.60%, Background 77.80%, Challenge 46.60%, Future Work 67.50%, Outcome 67.90%. Lauscher et al. (2018) uses this annotated data but takes a different approach using a neural model based on a recurrent neural network with long short-term memory cells (LSTM) to predict labels. They only show F1 % scores for each annotation schema, not individual labels. The neural network approach produced low performance: Scientific Discourse 42.70%, Subjective Aspect 18.80%, Summary Relevance 33.50%. It is likely that the limited amount of training data, only 40 papers, available to the neural network results in these low scores.

2.2.2 Citation Function Models

A *Related Work* section will, by its nature, include citations to other work and reference to the field in general. Understanding the motivations or function of a citation can help determine an author intention (Teufel et al., 2006b). Aspects of giving feedback to a writer will require an understanding of what function a citation undertakes, such as whether the citation is just descriptive or provides a comparison to the author's own work to signal a gap, or is used by the author to build their own work.

Work on citation function schemas has been an area of research for several decades, with more recent work considering how this recognition can be automated (Weinstock, 1971; Oppenheim and Renn, 1978; Teufel et al., 2006a; Angrosh et al., 2012). Previous work in AZ underpins some of the successful work in automated recognition of citation function, such as (Teufel et al., 2006a; Jurgens et al., 2018; Siddharthan and Teufel, 2007). Many schemas for citation function do not label at a sentence level as it it is known that authors will not criticise a paper outright, often they will use hedging or give praise, then in the next sentence mention a negative aspect (MacRoberts and MacRoberts, 1984). Sentence labelling can fail to capture the overall intention of the author in citing the work and understanding the citation context window is vital in capturing correct citation function (Ritchie et al., 2006, 2008).

Whilst many aspects of citation function works are related to this work the most meaningful is probably that of Angrosh et al. (2012) carried out on *Related Work* sections, and that of Teufel et al. (2006a,b) done within Computational Linguistics – these works cover both the annotation and the automation of their schema. Both models are described below.

2.2.2.1 Citation Function Model (Teufel et al., 2006a,b)

This citation function schema in (Teufel et al., 2006b) is designed for information retrieval, such as improving citation indexing or enhancing bibliometric measures. It consists of 12 categories, as seen in Figure 2.5. Citations in 360 conference papers from Computational Linguistics were annotated, with each given a single category label. Three annotators annotate the articles, and Kappa is used to measure inter-annotator agreement on 26 articles, where Kappa = 72%.

Category	Description
Weak	Weakness of cited approach
CoCoGM	Contrast/Comparison in Goals or Methods(neutral)
CoCO-	Author's work is stated to be superior to cited work
CoCoR0	Contrast/Comparison in Results(neutral)
CoCOXY	Contrast between 3 cited methods
PBas	Author uses cited work as basis or starting point
PUse	Author use tools/algorithms/data/definitions
PModi	Author adapts or modifies tools/algorithms/data
PMot	This citation is positive about approach used or problem addressed (used to motivate work in current paper)
PSim	Author's work and cited work are similar
PSup	Author's work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

Table 2.5: The twelve labels of the citation function schema from (Teufel et al., 2006a) with a description of each label in the right hand column.

Automated classification is done with the IBk algorithm with 10-fold cross validation. Overall accuracy reaches 77% and MacroF 57%. The scores for individual labels show variation again with sparse labels having lower scores. Scores range from 1% to 62.70%. Collapsing categories into three high level citation functions, seen in Figure 2.6, increases classifier performance accuracy to 83% and Macro-F 71%. When the distribution of labels is highly skewed, classifier performance often improves by aggregating labels under a few coarse-grained categories, as Teufel et al. (2006b) has shown.

Old Categories	New Category
Weak, CoCo-	Negative
PMot, PUse, PBas, PModi ,PSim, PSup	Positive
CoCOGM, CoCoRO, CoCoXY, Neut	Neutral

Table 2.6: How the original categories of labels are collapsed into three citation function labels in (Teufel et al., 2006a).

Label	F1 % Scores	Label	F1 % Scores
BG	93	RWD_CS	97
RWSC	94	RWD	92
CWO	94	RWS	59
CWOW	83	ASRW	60
CWSC	54	RWO	49
AWRW-CS	33	RW_CW	-
RWO_CS	22		

Table 2.7: F1 scores for label prediction from Angrosh et al. (2012).

2.2.2.2 Citation Function Model (Angrosh et al., 2012)

The schema focuses on context identification designed to support information retrieval, supporting links between research papers and researchers for information needs, such as intellectual lineage, who else works in this area, similar approaches. 13 labels are applied at a sentence level (Table 2.8). *Related Work* sections from 50 research articles of Lecture notes in Computer Science are used. No information or results are provided of the annotation study. The model is trained using CRF and achieves very high accuracy of 93.22%. F-Scores for individual labels can be found in Table 2.7. Whilst these results seem remarkable, there is one thing to note that it is an unbalanced data-set. The five labels achieving over 90% F1 Scores represent 95% of the data; any labels below 90% F1 score are very sparse within the data set. Liakata et al. (2012) uses CRF when predicting Core Scientific Concepts but yet this performed worse than the SVM model, so why does this work perform so much better on such a smaller data-set? This is most likely to do with the feature set. The features are based on terms that are classified into categories being present in a sentence or in a previous sentence, and on whether citations are present or not in a sentence or a preceding sentence. Features are defined in terms of specific lexical items or citations. Angrosh et al. (2012) do not say how they have identified lexical items. If they come from the same data set they have used for labelling, the high accuracy they achieve may be a result of over-fitting.

Angrosh Labels

Class	Label	Description
Background/Intro Sentences		
Background	BG	Background Sentence describing background in research area
Citation Sentences		
Rel Work Desc - Citation	RWD_CS	Citation sentence describing the related work
Rel Work Outcome - Citation	RWO_CS	Citation sentence pointing out an outcome of the related work
Rel Work Strengths - Citation	RWS_CS	Citation sentence describing the strengths of the related work
Descriptive Sentences		
Rel Work Desc	RWD	sentence describing the related work
Rel Work Outcome	RWO	sentence pointing out an outcome of the related work
Rel Work Strengths	RWS	sentence describing the strengths of the related work
Research Gap Sentences		
Related Work Shortcoming	RWSC	Sentence noting the shortcomings in the related work
Contrasting Work for a Related Work	CWRW	Sentence describing contrasting work for a related work
Alternate Approaches		
Alternate Work for a related work citation sentence	AWRW-CS	Citation sentence pointing out an alternate work for a related work
Alternate Work for a related work	AWRW	Sentence pointing out an alternate work for a related work
Current Work		
Current Work Outcome	CWO	Sentence describing the outcome of the current work
Current Work Shortcoming	CWSC	Sentence describing the shortcomings in the current work

Table 2.8: Labels for sentences in a *Related Work* section from (Angrosh et al., 2012) with longer label names in the left column, the shorthand for a label in the middle column and the description for each label in the right hand column.

Argumentation Theory

Type	Description
Claim:	A statement that something is so.
Data:	The backing for the claim.
Warrant:	The link between the claim and the grounds.
Backing:	Support for the warrant.
Modality:	The degree of certainty employed in offering the argument.
Rebuttal:	Exceptions to the initial claim.

Table 2.9: Argument types from the Toulmin Model of Argumentation theory in the left column and their description in the right column.

2.2.3 Argumentation Theory Mining

Argument mining aims to identify the relevant components of an argument automatically (Peldszus and Stede, 2015). Argument mining has been applied to areas such as opinion mining, summarisation, and automatic essay scoring (Egan et al., 2016; Ghosh et al., 2016; Barker and Gaizauskas, 2016). These works consider aspects of argumentation theory based on the Toulmin model of argumentation (Toulmin, 2003). This model is a tool to analyse or represent arguments made by a writer (or speaker) and consists of recognising elements within an argument. The model is described in Table 2.9.

Recently, elements from this model have successfully been used to represent arguments within persuasive essays and to predict the quality of these essays. The idea is that the argument structure would correlate with essay quality. Works such as (Song et al., 2014; Ghosh et al., 2016) have shown that incorporating elements from argument mining to predict essay scores can be matched to human ratings. This is not dissimilar to the intuition in this thesis, that author intentions could be used to predict quality, as seen in Chapter 8. Nguyen and Litman (2018) point out, however, that most studies in argument essay scoring are limited in that they are not fully automated to extract features, and do in some instances rely on hand-marked annotations demonstrating the difficulty of the task. Their work investigates how argument mining can add value to essay scoring, rather than just further improving the state-of-the-art, suggesting it can extend and add

additional information, reasoning about the argument rather than statistical and lexical features related to the score. This idea also adds value to writing support as knowledge about the argument structure and what is missing, or present could enable better feedback and suggestions for improvement.

The majority of the work which has a basis in argument mining theory, however, is carried out on persuasive essays rather than research articles. Persuasive essays differ from research articles in that they focus on the writer presenting their views on a topic, while a research article contains aspects of this; it is generally a more extended piece based on much more in-depth research presenting facts and conclusions established through experimentation. This model of argument theory representation may be suited to research articles that are more theoretical or discursive of theories but unlikely to work well for providing feedback on writing within a scientific discipline.

Kirschner et al. (2015) develop an annotation scheme built on argument theory for articles within the education domain. This model considers relations between sentences, proposing four binary relations: support and attack, taken from the argumentation model; detail, which roughly corresponds to background or elaboration; sequence, as described in Rhetorical Structure Theory (RST) (Mann and Thompson, 1988); presentation, e.g. firstly, or subject matter, e.g. before. They suggest that models like *Argument Zoning* and *CoreSC* are too coarse-grained and only reflect the standardised way in which papers are written, not revealing how an author connects their thoughts to construct an argument. Whilst this later point by Kirschner – that these models do not take sentence connections into account – is true about AZ and CoreSC, it is a point Teufel herself raises. She highlights an understanding of sentence relationship and looking beyond sentence information may support better labelling (Teufel and Kan, 2009). The authors of Research Writing Tutor also raise this, suggesting it will be the next enhancement they can bring to improve accuracy in predicting sentence labels (cf. Section 2.3.1.1). The work in this thesis does implement aspects that draw on context and features beyond a single sentence to label author intentions, described in Section 5.3.

There are other areas of work in argument mining that have relevance, although they cannot be directly applied in our case of writing support to the task of labelling sentences with intention. For example, research within argument reason-

ing comprehension, such as the *SemEval Argument Reasoning Comprehension Task* (Habernal et al., 2017, 2018), or the use of argumentation in fact extraction and verification in the *SemEval FEVER Task* (Thorne et al., 2018b,a). The Argument Reasoning task focuses on the idea of comprehending an argument through identifying and reconstructing warrants, but this is a complex task as often warrants are implicit and require the reader to make an inference. The idea of implicit warrants is relevant to our task of discovering author intentions, as it is known that writers of academic text often leave the reader to infer contextually. Later in this thesis, we see evidence of where annotators make such contextual inferences (cf. Section 4.8.3). Our machine learning approach struggles to label the sentences where this occurs, because there are no surface indicators (cf. Section 5.8.2). Similar to our observations in this thesis, the task organisers, when analysing the results from the competing entries, found that discriminating surface features helped, but failed when these were misleading. In addition, they found that the inclusion of external knowledge is key, with strong entries relying on 'inference corpora' in pre-training steps for models.

The FEVER task looks to validate textual claims from textual sources. This task is somewhat different from ours, where we only look at the writing in a *Related Work*, they look to verify claims using information retrieved from a large set of external documents. This type of method could be used to compare author intention statements about cited works and corroborate with evidence in the cited document source, leading to more robust labelling of intentions. Analysing the FEVER task entries, the organisers also found similar evidence to the reasoning task, namely that pre-trained models on natural language inference produce better performance.

2.3 Automated Writing Evaluation

Automated writing evaluation (AWE) has been an area of research dating back to the 1960s (Hockly, 2019). Research in this area utilises technology such as NLP, statistics and machine learning to understand aspects of style, grammar, complexity, topics, and to evaluate and score written discourse. Popular commercial tools include Criterion[4] developed by Education Testing Services (ETS), Write

[4]http://www.ets.org/criterion/

& Improve[5] from Cambridge English and MyAccess![6] from Vantage learning. The use of AWE tools in large-scale assessments within education systems has also become more prevalent in the last two decades with e-rater, now part of Criterion, deployed in 1999 to carry out operational scoring of Graduate Management Admission Testing (GMAT) of analytical writing assessment. GMAT have subsequently replaced this with IntelliMetric[7]. IntelliMetric expands on scoring to allow for student revision and editing, and to provide feedback on rhetorical and sentence level dimensions (Cotos, 2014, Ch. 2).

Much of the focus of work within AWE is on school education or undergraduate-level essay writing, not on post-graduate academic writing, except for a small number of works, such as Research Writing Tutor, AcaWriter, Mover (Cotos and Pendar, 2016; Abel et al., 2018; Anthony and V. Lashkia, 2003) (we describe several of these tools in more detail in this section). The tools applied to writing evaluation, and not providing essay scores alone, are all based on hand-crafted feature approaches. Hussein et al. (2019) in their survey of automated language essay systems highlight that none of the tools that provide feedback use neural approaches, and all rely on hand-crafted features. One reason we may see less focus on neural approaches for formative evaluation of writing is the need to understand and validate the outcomes from such systems. Hand-crafted features relate closely to the rubrics that are employed by individuals during assessment (Hussein et al., 2019). On the other hand, neural approaches are more difficult to interpret and to understand as to what features are learned. There is no way to know if the features learned are those used by human raters to predict a score, or if the neural approach is learning idiosyncrasies of the raters themselves (Hussein et al., 2019). Another reason is that, until the recent introduction of neural approaches based on pre-trained models (e.g. BERT (Devlin et al., 2018), ELMo (Peters et al., 2018)) these models needed large amounts of annotated data. Whilst there are large data sets of essays with matching scores, gaining similar sized data-sets for the annotated critique of writing is time consuming and expensive.

More recently though, we see neural approaches used in developing state-of-the-art for essay scoring, although many results highlight that mixing both hand-

[5] https://writeandimprove.com
[6] http://www.vantagelearning.com
[7] http://www.intellimetric.com/direct/

crafted features and neural approaches improves performance. Nadeem et al. (2019) show that for scoring persuasive essays, a feature-based system outperforms the neural model approach. In further work with their system, it is argued that combining the neural and the feature-based approach should be exploited, as this provides the state-of-the-art performance (Liu et al., 2019a). So, whilst neural approaches are capable of deriving features automatically, the strength of feature engineering is that it is more interpretable (Jiang et al., 2018; Nadeem et al., 2019). The ability to interpret and understand how features relate to author intentions is vital in our work of providing feedback, and may give us valuable insight into pedagogical aspects. We see evidence of this in Section 5.8.2.

2.3.1 Automated Writing Evaluation Tools

In this section we describe writing tools that are most closely related to the work carried out in this thesis.

2.3.1.1 Research Writing Tutor

Research Writing Tutor[8](RWT) described in (Cotos and Pendar, 2016) is designed to label sentences within an *Introduction* section as communicative moves and rhetorical steps based on Swales CARS model. It consists of two support vector machine (SVM) classifiers that cascade to predict the moves and the underlying steps associated with each move. The underlying data that is it built on is *Introduction* sections from Journals spread across 51 disciplines with 20 articles from each. Annotation was performed by three annotators, but the authors did not indicate how many papers were double annotated. Annotation agreement was done following calibration meetings across 30 texts from different disciplines, reaching high agreement measured with Intra-class-correlation (Moves, r=0.86, Steps, r=0.80). Training for the classifier was done on 650 *Introduction* sections. Features mainly consisted of uni and tri-grams. The results for the two step classifier are good with an overall accuracy of 72.60% for moves and 72.90% for steps. Like most classification systems where moves or steps have sparser representation scores are lower. **Move 2 - Establishing a niche** has less data available with only 926 sentences and only achieves an F1 Score of

[8]https://cce.grad-college.iastate.edu/resources/writing-resources

45.80% compared to **Move 1 - Establishing a Territory**, which has 3,233 sentences and reaches a F1 score of 80.40%. The authors themselves acknowledge the approach reaches its limitations with only using n-grams. This restricted feature set, within one sentence, limits the ability to take context into consideration or any sequence information that may provide more discriminant classification, aligning to the points made earlier in Section 2.2.3. What is not specified is the actual disciplines the data is from or any indication as to how diverse or similar these may be. Whilst the authors' claim the work is discipline independent, there is no analysis provided to show to what extent this is true. Given the evidence of linguistic variation across disciplines and the relatively little number of papers used from each discipline, more information is needed to support this claim. This tool is not available outside the University faculty it is being developed within.

2.3.1.2 ACAWriter

ACAWriter[9] is another example of a writing analytic tool. This tool has gone through several iterations and the description here is based on their latest paper (Abel et al., 2018). The tool is based on recognising rhetorical moves in *Introductions* and *Abstracts* although other works have looked at extensions for reflective writing and essays support for under-graduate law students (Gibson et al., 2017; Knight et al., 2018). The parser engine is based on a concept-matching framework described in Sándor et al. (2006). This parser operates at a sentence level and ties together both matching expressions, e.g. cue phrases that convey concepts but also considers grammar dependencies within a sentence in order to classify rhetorical moves. The original tool classified into eight rhetorical moves, described in Table 2.10. When these eight rhetorical moves were mapped to Swales CARS model two were dropped, Trend(T) and Surprise(S). Table 2.11 shows the mapping of AcaWriter's tags to the CARS model. The authors fail to include in the paper any metrics on the accuracy of their parser, so its performance cannot be compared against the current work. The development of AcaWriter is firmly rooted in the Learning Analytics field, and as such, there is a focus on the effective design and presentation of author intentions for student feedback. AcaWriter motivates the way in which our writing analytic tool **LitCrit** is designed, described in Section 6.2.

[9] https://www.uts.edu.au/research-and-teaching/teaching-and-research-integration/acawriter

Rhetorical Move	Tag	Example
Question	Q	Current data is insufficient to conclude that ....
Background	B	Recent studies indicate that ....
Contrast	C	In contrast with previous hypotheses
Emphasis	E	Studies on x have provided important advances
Novelty	N	This model provides a new approach to ...
Surprise	S	This discovery of x suggests intriguing ...
Trend	T	New Models of x are emerging ...
Summary	S	In this paper we show that ...

Table 2.10: AcaWriter's rhetorical move labels in the left column with their shorthand notation in the middle column and an example sentence in the third column taken from (Abel et al., 2018).

CARS Rhetorical Move	AcaWriterTag
Move 1 - Establishing a Research Territory	E - Emphasis, B - Background
Move 2 - Establishing a Niche	C - Contrast, Q - Question
Move 3 - Occupying the niche	S - Summary, N - Novelty

Table 2.11: The mapping of the CARS model moves and AcaWriter's rhetorical tags from (Abel et al., 2018).

Criterion Discourse Labels

Label	Description
Conclusion	Segments that summarise the essays entire argument.
Introductory Material	Segments provide the context or set the stage in which the thesis, a main idea, or the conclusion is to be interpreted.
Main Points	Segments assert the author's main message in conjunction with the thesis.
Support	Segments provide evidence and support the claims made in the main idea, thesis statement, or conclusions.
Thesis	Segments state the writer's position statement and are related to the essay prompt.
Other	Segments not following into the above categories.

Table 2.12: The six discourse labels and their descriptions used in Burstein et al. (2003).

2.3.1.3 Criterion

Criterion[10] is a commercially available product providing feedback for high school essays designed for specific disciplines. The tool has undergone several iterations and here the work most pertinent to this thesis that of discourse labelling is described based on (Burstein et al., 2003). The discourse labels are designed to help a student think about the organisation and development of their writing by highlight the text with the discourse labels and highlight aspects that are missing. Discourse labelling is based on 6 labels in Table 2.12.

Annotation was carried out in iterative phases with inter-annotator agreement tested at multiple stages, and Kappa values were kept consistently above 80%. The final number of annotated essays used in classification is 1,462 essays. The classification engine is based on a voting algorithm that takes the decisions of three independent classifiers, one decision based and two probabilistic. It clas-

[10] https://www.ets.org/criterion

sifies on several features including rhetorical relations, discourse marker words, terms, cue phrases, syntax and sentence mechanics, e.g. punctuation, sentence number, paragraph number. Their voting system shows the best performance over using the classifiers singularly, with overall F1 score of 85%, individual F1 scores: Introductory 57%, Conclusion 84%, Main point 77%, Other 76%, Support 91%, Thesis 73%. In addition, they show essay topic independence between the six topics within the data set with accuracy for each topic F1 scores ranging from 74% to 82%. These results are very encouraging for a writing feedback system, but it would be useful to know what the distribution of the labels was within the data set to ascertain if those that performed better were more abundant, as is observed in other results discussed in this chapter. A strong factor in the performance will be the high number of annotated sentences as will the relatively small number of labels. One aspect of the data that is not discussed is what quality the essays used are. If they are all of high quality they will more likely contain a higher number of training examples of all label types.

2.4 Other Aspects of Writing Beyond Author Intentions

Whilst novice writers may struggle to form the correct author intention and meet the function of a *Related Work* it is also possible they may have difficulty expressing this coherently, meaning automated feedback may have to deal with general aspects of poor quality writing. Many people have written about academic writing, the challenges, and how to succeed in writing good quality articles. The focus of this thesis is only on the author intentions and not on correct grammar, spelling or creative style. However, the importance of fluency and readability of the text should not be under-estimated as this has a strong bearing on the ability to extract meaningful content for humans, but it is likely to be even more challenging for an automated system. Early works (Kincaid et al., 1975) developed metrics for predicting reading difficulty levels using characteristics, such as sentence length, characters in words, numbers of syllables per word. More recent works include linguistic features such as word frequency incorporated into language models (Si and Callan, 2001; Collins-Thompson and Callan, 2005) or syntactic features (Schwarm and Ostendorf, 2005). However,

as these aspects are deemed out of scope of this research, the approach to dealing with this is that only published articles are used in experiments. It is assumed that these articles reach a minimum acceptable threshold within their domain. However, in our conclusions, we do acknowledge that such aspects may need to be included in future possibly utilising existing tools, such as Grammarly[11] or Turnitin[12].

Another aspect that novice writers may struggle with, not related to author intentions, is in the structure and presentation of the writing. This should flow coherently across topics and guide the reader with appropriate use of discourse relations to signal direction and transition smoothly. Work has focused on discourse relations and entities or anaphora resolution and how intentional use of these support coherence. Several models have been proven effective in studying these entity shifts such as centering theory (Grosz et al., 1995) and rhetorical structure theory (RST) (Mann and Thompson, 1988). Work specifically relating coherence to essay scoring was carried out by Miltsakaki and Kukich (2004) who capture a source of incoherence that they link to lower essay scores through considering rough shift patterns in entity transition, based on centering theory. These signals of relations have been shown to provide cohesion and coherence within a text linking to the quality and well-written nature of a text (Pitler and Nenkova, 2008) and that there are orders that are favoured that provide coherence to the reader (Louis and Nenkova, 2012). In the user study in Section 3.2, evidence is found that PG students are more likely to notice that intentions are missing in works that have rough entity shifts. However, presentation and structure is not an aspect pursued in this thesis.

2.5 *Related Work* Section Writing

There are many resources of books, web help pages, university courses dedicated to providing information and guidance on writing about literature work, but PG students still struggle in this area (Boote and Beile, 2016). Kamler and Thomson (2006) in their book, which focuses on helping PhD students write, provide relevant insight into the problem. They argue that there is a lack of focus on writing

[11] http://www.grammarly.com
[12] https://www.turnitin.com

in context treating it instead as several discrete skills that are not contextualised, focusing on writing as skill-based steps that once learned, can be applied. In contrast to this, they argue that *research is writing* and writing is a continual part of the process of research, embedded within it. Perhaps though what is of most interest for a *Related Work* section is what they say about why PG students struggle when talking about others' work. They highlight that students struggle the most with being critical, where they must offer an opinion and take a stance about others' work. However, the novice is new to the field and perhaps unaware of all the histories of debate and concerned that the authors' of these cited works may become their reviewers or thesis examiners. The novice writer must learn to develop their voice in the sea of more esteemed and learned colleagues. They question the novice writer's true understanding of what being critical means, in that it is more than just praising or being critical but understanding what work to include, ignore or how to adopt a critical stance and bring perspectives together to establish aspects of similarity or differences and consider how the work contributes. Evidence to support this is found in the next chapter when reviewing experts' and students' opinions on *Related Work* sections. The results show that PG students struggle with the concept of being critical and do not yet seem to have developed a full understanding, with most focusing on criticality as a means to offer merits or drawbacks only but not understanding the importance of this in the discussion.

Although in this work, a study is carried out with experts to develop an understanding of what makes a good *Related Work*, there is a need to understand what the literature says about *Related Work* sections in order to support the design of the study. There are many resources available, however, the most informative and useful description of what makes a good *Related Work* that we have come across is that of Kamler and Thomson (2006). They describe the key tasks that literature work should accomplish:

1. Sketch out the nature of the field or fields relevant to the inquiry, possibly indicating something of historical development and
2. identifying major debates and define contentious terms, in order to
3. establish what studies, ideas and/or methods are most pertinent to the study and
4. locate gaps in the field, in order to

5. create the warrant for the study in question, and
6. identify the contribution the study will make.

It is this explanation that lays the foundation for the design of the study in the following chapter.

There are differences between a *Related Work* section and a *Literature Review* section in a thesis pertaining to length and depth but none the less a *Related Work* section is still expected to serve a purpose. However, not all articles are expected to contain a *Related Work* section, and this can be discipline specific, or it can be related to the venue the paper is appearing in, such as workshops or papers for poster submissions which tend to be short, e.g. four or fewer pages. This though does not make the content of a *Related Work* section irrelevant or not useful for a PG student to learn as there is still an expectation that this will be found somewhere in the paper, rather than in one concentrated section.

2.6 Insights in the Decline of *Related Work* Writing

A recent interesting study (Jurgens et al., 2018) carried out an investigation of the evolution of citation framing across the ACL anthology paper, a field focused on NLP. This study provides a different perspective on why there may be a decline in how researchers talk about related work. They look at citations across a whole paper but the results are still relevant when thinking about a *Related Work* section and how and when styles may differ when an author talks about other's work in relation to their own. Their analysis of citation framing reveals notable relationships between writers, readers and the discipline and interesting changes as a field develops. Firstly, they show that the way citations are framed is impacted by the venue a paper is published in. For example, journals have the highest percentage of background citations; conferences give considerably more space to contrast and comparison to other work; workshops have relatively little comparison and more focus on background citations or citations that show how the author uses other work. This, however, changes as a workshop matures and it becomes more conference like in its citation framing. Secondly, they show that certain types of citation framing are significantly predictive of a paper having a higher impact. This is true of two types of citations (i) when an author employs

framing they describe as **USES** – an author cites to show they use other work (ii) when an author frames their contribution through **COMPARISON** or **CONTRAST** framing against other people's work. Finally, they show how citation framing as a whole shifts as the NLP field has matured. There is a reduction in framing that is needed for positioning and excessive comparison, with an overall shift towards rapid science discovery.

This last finding is one of the most interesting in that it raises further questions. Previously mentioned is the struggle that PG students have when writing about others' work. It is possible that novices are perpetuating this lack of positioning or comparison to other work as they see this as the norm in the materials they read. It is hard to know exactly what impact this is having on novices but it seems plausible that what one reads influences their own writing. It would also be interesting to see if a similar study across another discipline would observe similar effects.

2.7 Summary Discussion

This chapter has highlighted related work from several different fields that are relevant to our goal of helping PG students with writing their *Related Work* section. In this section, we summarise and show how gaps in current work relate to our contributions, and describe our hypotheses and our approach to filling these gaps. We break this down into four areas: building a model of author intentions specifically for *Related Work*; building an effective feedback tool for writing support; recognising author intentions automatically; investigating if the model can apply across disciplines.

2.7.1 Building an Author Intention Model to Support *Related Work*

All the models discussed in this chapter are designed to solve a specific informational need, and the model labels proposed reflect this. Additionally, the models are built for specific sections, e.g. the *Introduction* or for a whole document, and can also have intentions that are not relevant, e.g. Conclusions, Aim, or are missing intentions relevant to a *Related Work*. Additionally, individual labels of

models can appear similar, but they can still differ in application, meaning they do not convey the same information. These individual labels of a schema directly relate to the content expected to be present by a domain expert, within the context of the informational need. There currently exists no model of author intentions to support writing feedback in a *Related Work* Section. It is, therefore, a requirement within this thesis to understand what content should be present within a *Related Work* section in order to build a model of relevant intentions for writing feedback within this section.

The approach to building author intention models in current research has been from observations of domain experts through studying scientific papers (cf. Section 2.2.1), or in the field of AWE where the rubrics set out by domain experts for marking are used to help define a schema. However, there is no evidence that these characteristics from observational studies are the same as the characteristics that experts use when conducting peer-review. Whilst there are considerable resources available to undertake the same approach, we believe that using peer-review is a valid alternative for obtaining an understanding of the content expected to be present. One possible problem though is to understand how much experts agree, as disagreement during peer-review is a widely known problem (Lawrence and Cortes, 2014). We hypothesise that if we can find a reasonable agreement between experts about content expectations during peer-review, then this can be shown to be a valid approach to developing an author intention model.

Another advantage of using peer-review is that it will allow us to compare PG students' responses to expert responses. Novices are known to have difficulty differentiating good work from bad and can lack the perspective to identify important points (Cambre et al., 2018). We hypothesise that the comparison of the experts' and PG students' peer-review will allow for a greater understanding of where PG students struggle. An understanding of how PG students differ will help focus the feedback generated from our **LitCrit** tool on aspects that will benefit PG students more.

Our discussion of existing automated writing tools that provide feedback and scores shows that there is a relationship between author intentions and quality, and intentions can contribute to automated scoring. The author intention model we build in this work represents the content expected to be present in a *Related Work*. We hypothesise that these intentions should represent a quality measure

of a *Related Work*, and we test this by using the intentions to predict the quality of a *Related Work* section.

The research questions explored in supporting the hypotheses above are:

1. What are the content expectations highlighted in a *Related Work* by experts and do experts agree with each other?
2. Do PG students differ from experts in what they look for in a *Related Work*?
3. Do the author intention labels serve as a proxy for indicating content quality in a *Related Work*?

2.7.2 An Effective *Related Work* Feedback Tool

In this chapter, we described several writing tools that do employ the use of author intentions in giving feedback. These tools work by highlighting narrative to bring aspects to the attention of the reader. It is argued that drawing the attention of the reader will influence their thinking, helping them to develop their writing. However, there are no studies that evidence this change of thinking. We hypothesise that such a difference in thinking can be measured by comparing PG student responses during peer-review with and without this highlighting of the narrative.

The research questions explored in supporting our hypothesis is:

1. Does highlighting the narrative structure with intentions change PG student perceptions of a *Related Work*?
2. Do PG students find the visualisation of intentions and feedback on missing aspects helpful?

2.7.3 Approach to Automating Recognition of Author Intention in a *Related Work*

Results found for existing methods for predicting author intentions in academic writing are promising, with some achieving levels of accuracy that would be acceptable for writing feedback, e.g. greater than 70% accuracy. However, as always, sparse categories cause problems for prediction accuracy. Although many

recent advances in NLP have been facilitated by neural network approaches, we have not seen their use yet in automated writing evaluation. One of the reasons for this is that the education research community places a greater value on the ability to interpret and understand (1) the meaning of features with respect to outcomes and (2) their alignment with a marking rubric, than they do on overall system performance. However, we do see (cf. Section 2.3) that the introduction of neural approaches within automated essay scoring does improve the state-of-the-art performance, but performance remains optimal when handcrafted features are incorporated into a neural approach. We did see the use of a neural model in classifying author intention models in ArguminSci (Lauscher et al., 2018) (cf. 2.2.1.4). The results produced though would not have been good enough for writing feedback, likely due to the small, labelled data-set.

Traditional feature engineered approaches rely on using n-grams, which can create large sparse representations or approaches that use lexicons which create a more compact representation but suffer from out-of-vocabulary words. Word embedding approaches use pre-trained word embeddings rather than n-grams as a feature (Mikolov et al., 2013; Pennington et al., 2014). Embeddings translate a word to an embedding vector making it possible to model the semantic importance of a word in a numeric form and, thus, perform mathematical operations on it. These embeddings have the ability to generalise better by capturing semantic or syntactic information. However, initial methods were sub-optimal, as words can have multiple senses and these types of embeddings only allow for a single representation. This has led to the introduction of contextualised word embeddings, e.g. ELMo (embeddings from language models) (Peters et al., 2018).

Neural models often need large amounts of labelled data to train from, and prelabelled data is expensive and difficult to acquire due to expert annotators being needed. Word embeddings provide access to unsupervised pre-training of large corpora resolving this need for annotated data, and have led to significant advances in state-of-the-art in NLP. This has been further advanced by developments at Google – including a novel neural network architecture which uses *transformers* (Vaswani et al., 2017). Transformer neural architectures have many benefits over other neural approaches, such as providing more effective ways to model long term dependencies in temporal sequences, and eliminating the sequential dependency on previous tokens through more efficient training. Google

also introduced BERT (Bi-Directional encoder representations from transformers) (Devlin et al., 2018). Unlike other models which use unidirectional language models to learn, BERT uses a bidirectional language representation and this implementation allowed them to show significant advances in the state-of-the-art in NLP tasks (Devlin et al., 2018).

Whilst embeddings are now a dominant approach in NLP there are several aspects that must be considered. Firstly, these contextualised embeddings are trained on general domain corpora, e.g. Wikipedia or news articles, and as highlighted in this chapter language used within the scientific domain not only differs to that of general corpora, but also differs between disciplines. We highlighted earlier in this chapter that tasks in argument recognition benefited from being pre-trained on more subject appropriate corpora. SciBERT (Beltagy et al., 2019), released late last year, is modelled on BERT but uses data from Semantic Scholar to pre-train. They showed that the vocabulary in BERT compared to that in SciBERT only had a 42% overlap. Using SciBERT in the task of classifying citation functions from (Jurgens et al., 2018) Beltagy et al. (2019) demonstrated overall a 17.98% increase in F1 score compared to (Jurgens et al., 2018) original result, which used a machine learning approach with engineered features. The second aspect to consider, as in our case, pre-trained embeddings do not exist or do not adequately match our domain. Creating such embeddings is extremely compute-resource and time expensive, thus, not always feasible to achieve.

Decisions on the approach to take in order to automate the classification of author intentions in this thesis were taken considerably before the adoption of BERT in NLP tasks. However, had they been taken later, we would have no doubt pursued or undertaken a comparative study particularly with the availability of SciBERT. However, this thesis is an exploratory look at author intentions to automate feedback on a *Related Work* section; thus, the focus is not necessarily on the best performance but the ability to interpret and explain what the model is doing is very important. Being able to understand why a model generates errors will help to improve any areas of weakness and may give insights into any pedagogical gaps that might exist, as well as support more specific and more useful writing feedback. In addition, there does exist research that shows approaches using hand-crafted features with supervised approaches, such as SVM, can still outperform neural approaches (cf. Section 5.4). Therefore, a neural approach would

not necessarily be a guarantee for better overall performance. We hypothesise that an approach based on feature engineering will produce results that can support recognising author intentions. However, more importantly, this approach will allow us to explore errors and understand aspects important for feedback, or that relate to a better understanding of pedagogical implications.

Any approach, be it feature-based or neural, however, does require high quality annotated data which can be difficult, time-consuming and expensive to produce. Challenges were highlighted in this chapter how annotation with linguistically unmarked context brought subjectivity into annotation, leading experts to infer knowledge rather than annotate based on what is linguistically present. This demonstrates the need to carry out an annotation study to show that our proposed author model can be annotated with reasonable human agreement.

The research questions explored in supporting our hypotheses are:

1. Can the author intention labels be annotated with reasonable human agreement?
2. Can the author intention labels be recognised automatically with reasonable accuracy?

2.7.4 Discipline Independence of Approach Proposed

Author intention models discussed in this chapter were shown to require adaptation in order to be used in another academic discipline. Also observed were models that were built for specific sections, e.g. the *Introduction* or for a whole document, would have intentions that were not relevant, e.g. Conclusions, Aim, or were missing intentions relevant to a *Related Work*. Swales pointed out phrases that aligned to intentions, but other studies have shown that while linguistic patterns exist more often than expected by chance, they are subject to variation between disciplines (Biber, 2006; Biber et al., 2004; Cortes, 2004). Hyland (2008) reports that not only are such frequent word/phrase bundles central to academic discourse, but as they can be shown to differ between disciplines, they offer a means of differentiating between academic discipline writing. As described above, we take a feature engineering approach in this thesis, and hand-crafted features are often bounded by a domain (Hussein et al., 2019). This means our task of finding key phrases that align with our author intentions in one discipline

may not necessarily align to key phrases used within another discipline. Alternative approaches, such as those that use pre-trained models (Devlin et al., 2018; Peng et al., 2019) have led to state-of-the-art results as they develop the ability to generalise and build knowledge that can then be transferred. Using hand-crafted features could limit our ability to reach state-of-the-art in training our model to a new domain, and we discuss this further in Section 7.2

The majority of the work in this thesis is focused on one discipline, Computational Linguistics, but Chapter 7 considers how well the model of intentions and features for classification can be applied within another domain based on our feature engineering approach.

2.7.5 Summary of Thesis Contributions

In the discussion section we show the thesis covers four areas and these fall into two main contributions.

(1) We use peer-review to understand what content should be present in *Related Work* and build a model of author intentions to represent this content, using it to support writing feedback; and (2) we demonstrate that this author intention model can be reliably annotated by humans and build a classifier to reliably automate the recognition of these intentions within a *Related Work*.

Chapter 3

Understanding What Experts Look for in a *Related Work* and how PG Students Differ

3.1 Introduction

This chapter investigates through an exploratory study two main questions 1) What are the content expectations highlighted in a *Related Work* by experts and do experts agree with each other? 2) Do PG students differ from experts in what they look for in a *Related Work*? We report the findings from both groups and summarise this into a framework, highlighting where PG students struggle. The findings are used to build our author intention model in the next chapter.

3.2 User Study Design

The approach in this study is exploratory in order to understand what content experts look for in a *Related Work*, and if a consensus exists in the arguments experts expect to see in *Related Works*. Our approach is to use peer-review to study the content experts expect to be present. This is different from other works which have carried out observational studies to analyse what aspects are present. Therefore, no direct comparisons are possible between our work and existing research in terms of experiment design, but Gogolin and Stumm (2014), who

develop criteria and a framework to evaluate a publication through peer review, have relevant aspects. In designing our study, we start from their approach, adapting it to our specific task. Their study used a questionnaire based approach that assessed the demography of participants, and then multiple tasks to review papers with both closed questions, based on Likert responses, and open questions that invited the participant to provide a more comprehensive assessment of the paper under review. From their pilot study, it was shown to be better to have a smaller number of articles with more reviewers, in order to create more data for comparison; that scales for judging the articles should be longer (greater than 4); and that collecting demographic information was important to understand how heterogeneous the participant group may be with respect to their background. The use of open questions allowed the authors to gain a better understanding of the criteria the participants were using to make judgements on quality regarding the papers they were reviewing. Our experiment design is similar, using a questionnaire approach, and we also use the open questions to understand more about the criteria being used by participants to make judgements on the content of the *Related Work* sections. The closed question criteria used in (Gogolin and Stumm, 2014), however, are not specific enough to be used in understanding content that is expected in a *Related Work* section. They relate to aspects that are too general and found across a whole article. In designing our criteria for questions, we start from published sources on what should be present in a *Related Work* (Kamler and Thomson, 2006; Harmon and Gross, 2010) but also from (Boote and Beile, 2016). Boote and Beile (2016) undertook a study of PhD literature reviews and proposed a rubric by which to assess literature reviews. Literature Reviews differ from *Related Work* sections in an article, particularly by length and depth, so not all of the rubric criteria apply to our task. Maxwell (2006) in his commentary on (Boote and Beile, 2016) specifically highlights that their rubric does not address the relevance of the literature to the author's research. So, whilst this rubric provides a good starting reference, we do not use all criteria directly.

Also explored in the study is whether PG students highlight similar aspects, and what they struggle to notice compared to experts. Using peer-review allows for comparison of the expert group to the PG student group, which observational study would not. Research shows that novices undertaking peer-review often struggle to differentiate good work from bad, or to identify the characteristics

that matter (Sadler, 1989). Students also do not often get the opportunity to review work, other than their own, and have not had the opportunity to gather expert knowledge in making judgements or in understanding complex topics (Cambre et al., 2018). In addition, it is known that experts highlight deeper features during feedback, while novices can get misled by superficial features (Novick, 1988). Therefore, it seems a valuable medium to use peer-review to gain an understanding of where differences occur between experts and PG students when reviewing *Related Work*.

3.3 Methods

The study was set-up to answer our research questions and participants were requested to carry out three main tasks. Complete the consent and demographic questionnaire (cf. Section 3.3.2) , complete the opinions of *Related Work* (cf. Section 3.3.3) and to complete peer-reviews of seven *Related Works* (described in Section 3.3.4). This was all undertaken through an online interface. To minimise bias and fatigue, we undertook randomisation of the order in which *Related works* appeared to each participant, and we allowed participants to take four weeks to complete all tasks. This emulated a normal peer-review scenario, which allows a participant to return to the task over a period of time. In particular, participants were:

- Randomly assigned into four groups, each of which received the example sections in a different order
- Allowed to take a break as many times as needed
- Could navigate back to change responses or reread *Introductions* or *Related Works*
- Given four weeks to complete all tasks

The study received Ethics Committee approval from the School of Informatics. Examples of questions used are included in the main text in this Section, but a full set of questions can be found in Appendix B.

3.3.1 Study Participants

In recruiting participants, we take both a purposive and convenience approach to sampling (Kelly, 2009, Ch. 7, p. 67). The purposive nature of our sampling is that we only want to have experts or PG students from the field of Computational Linguistics. Our inclusion criteria for PG students is that they are currently undertaking a PhD or are a first-year post-doctoral. The inclusion criteria required experts to have first-authored ten papers and be post PhD for five years. This was considered a long enough period to have gained experience in writing *Related Work* sections, receiving feedback on the material they have written, and in reading other's work.

To recruit experts, an e-mail invitation to participate was sent to the author's network of academic contacts. Student participants were invited through University, School of Informatics mailing lists and the author's contacts.

Some practical limitations contributed to taking this approach, such as, the system of monetary compensation that was to be given to students, and the length of time that was needed to complete the experiment in full, which required some priming of participants to ensure they could commit. These methods of sampling fall into non-probability sampling, and can introduce bias, such as generalisability of results (Kelly, 2009, Ch. 7, p. 67). We recognise potential bias in our sample, in that the students and experts in our network are more likely to come from specific sub-fields in Computational Linguistics, as these are where the author's contacts reside. In addition, the students invited to take part will naturally fall into sub-fields within the School of Informatics Computational Linguistics groups. We see evidence for this bias in Figures 3.4 and 3.5 when we ask participants about what fields they have submitted papers to. We discuss this more at the end of this chapter in our section on limitations (cf. Section 3.8.1).

On recruiting the participants, they were sent an email which included a link to complete basic information to ascertain if the participant met the criteria. They were also informed that the examples of *Related Work* they would be asked to look at might vary in age and standard; thus they may not include all recent related works. They were not required to have an in-depth knowledge of the specific area and would not be expected to include a list of any missing relevant works.

All applicants met the pre-screening criteria, and subsequently all participants were then sent a link via e-mail inviting them to complete the study online. Students were given 15GBP compensation for taking part in the study.

3.3.2 Consent and Demographic Questionnaire

After consenting to participate, the participants were presented with a page that described the activities they would be asked to do, how to navigate the screens, save work and subsequently return. Next, a demographics questionnaire collected information on the participant's years of experience reviewing scientific literature, years of supervisory experience, year of first published paper, number of first-authored papers, the fields and application fields they have published in (taken from lists given to publishers when submitting at ACL conference), are they a native English speaker. If they were not a native English speaker, then questions about their academic education in English and years living/working in an English speaking country were asked. Additionally, they were asked to select an age category or to select *prefer not to say*.

3.3.3 Opinion Questionnaire about *Related Work*

After completing the demographic questionnaire, participants were then asked about their opinions on *Related Works*. This is considered a pre-task questionnaire which is commonly used to elicit perceptions and understand more about a participant's opinion or background knowledge (Kelly, 2009, Ch. 9, p. 91). Our opinion questionnaire is designed to gain an understanding in general of opinions of how important the participants think a *Related Work* section is, how they rate the importance of aspects and whether they think standards have in general declined. We did not find any previous questionnaires related to this specific topic, but Kelly (Kelly, 2009, Ch. 9, p. 165) investigates questionnaire mode and finds that subjects' responses when closed questions are used are significantly more positive when elicited electronically. Therefore, we provide a mix of questions types, both open and closed to elicit opinions from our participants.

The first two questions were free-text about a participant's opinion on the function of a *Related Work* and ask them if they look for any specific characteristics. The third question asks how likely they would be to reject a paper based on the

Related Work being inadequate or missing on a 4-point Likert scale (where *Very Likely* is 1, and *Highly Likely* is 4). We chose a 4-point scale to avoid a mid-point, which can sometimes be used when a participant does not have a strong opinion or does not know (Krosnick and Presser, 2010, Ch. 9, p. 271). In our case, some of the PG students may feel they have no experience in this area and thus no opinion. Questions 1-3 are shown in Figure 3.1.

The fourth question was multiple-choice about characteristics commonly thought to make a good *Related Work*. As described in Section 3.2 these questions were derived from the available literature that proposes what aspects should be present in a *Related Work* and from Boote and Beile (2016)'s rubric of what to look for in a literature review. The criteria we use is:(*Current Citations, Thoroughness, Context, Detail, Critical Evaluation and Extensiveness*. In this instance, we do want to provide a mid-point and thus use a 5-point Likert scale, where 1 is *Unimportant* and 5 is *Very Important*.

The next question asked the participant if there were any comments they would like to make about the characteristics and their importance in the previous question. The final questions were concerned with participants' thoughts on standards of *Related Works* in the last decade. The first, a multiple-choice question, asked if participants thought *Related Works* standards in the last decade had: *Got better, Declined, Stayed the same, Other (with a free text box to expand)*; if the participant thought standards had changed, they were asked if they could elaborate on why they thought this might be. Questions 4-6 are shown in Figure 3.2.

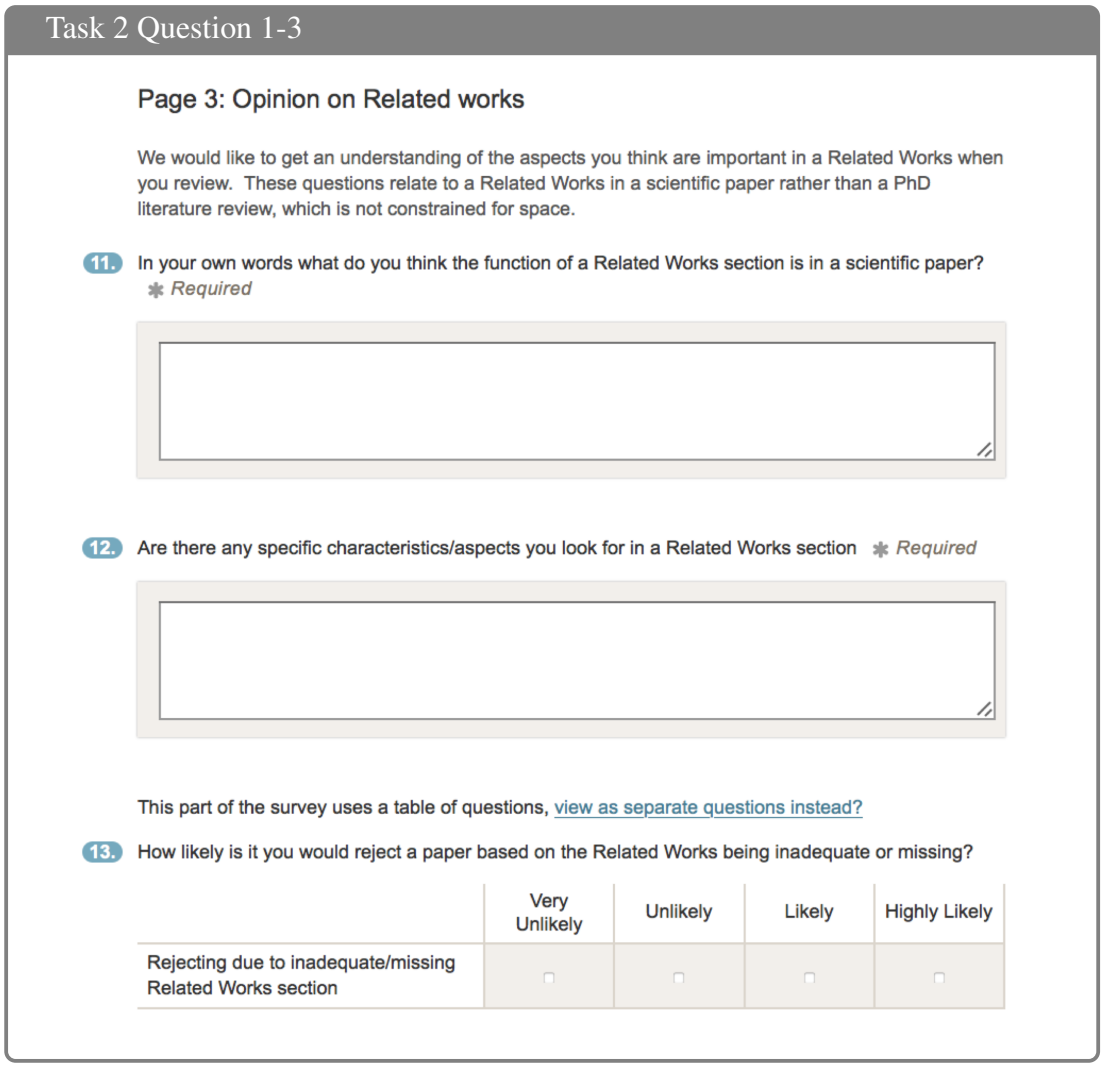
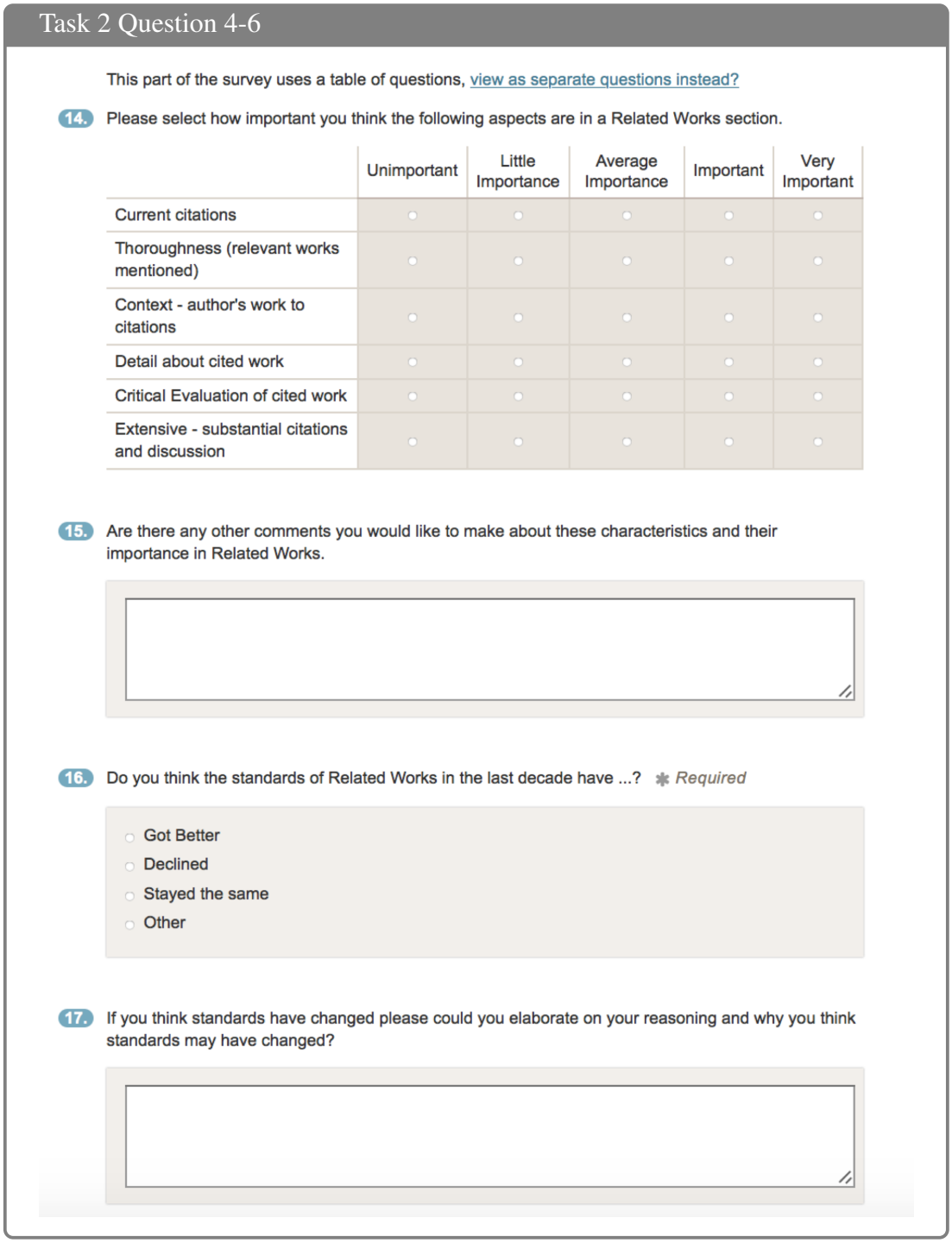


Figure 3.1: Questions 1-3 of Task 2 about the function of *Related Work* and likely rejection due to an inadequate *Related Work*



Task 2 Question 4-6

This part of the survey uses a table of questions, view as separate questions instead?

14. Please select how important you think the following aspects are in a Related Works section.

	Unimportant	Little Importance	Average Importance	Important	Very Important
Current citations	○	○	○	○	○
Thoroughness (relevant works mentioned)	○	○	○	○	○
Context - author's work to citations	○	○	○	○	○
Detail about cited work	○	○	○	○	○
Critical Evaluation of cited work	○	○	○	○	○
Extensive - substantial citations and discussion	○	○	○	○	○

15. Are there any other comments you would like to make about these characteristics and their importance in Related Works.

16. Do you think the standards of Related Works in the last decade have ...? * *Required*

- ○ Got Better
- ○ Declined
- ○ Stayed the same
- ○ Other

17. If you think standards have changed please could you elaborate on your reasoning and why you think standards may have changed?

Figure 3.2: Questions 4-6 of Task 2 about characteristics and the decline of *Related Work*

3.3.4 Main Task, Peer-review *Related Work* sections

3.3.4.1 *Related Work* Material Used

The number of papers to use was driven by the amount of time we believed would be feasible to ask of participants. We had estimated a reasonable time frame to be between three and four hours which equated to approximately seven papers allowing for two per hour. To select the seven papers we choose thirty papers from the data-set used in the experiment described in Section 8.3. These were papers published over the last ten years from the ACL anthology (Bird et al., 2008) which had been rated as poor, fair or good. We randomly selected ten of each rating, thirty in total. These papers were then reviewed by the thesis author and one supervisor to discuss how they could be manually modified to provide a range that emulated first draft stage to publishable quality, and seven final papers were chosen. Once the seven papers were selected edits were made to the content, but not grammar or spelling. For example, sentences discussing the author's own work or some comparison sentences may have been excluded or changes made to remove words such as *however* or *therefore*. There are alternatives to this approach in that we could have created our corpus by asking people to give us copies of their *Related Work* sections from previous drafts and final submissions. However, we believe that using published materials avoids issues of grammar or incoherence discussed in Section 2.4. In addition, we needed to ensure that the materials used covered a range of potential characteristics that could be missing or present when writing a *Related Work* (described later in this section). This of course introduces a potential bias of subjectivity from the author and their supervisor and may limit the generalisability of results (discussed more in Section 3.8.1). In addition, we recognise that the sample size is small, but fits within the constraints of the study.

Below a brief description of each paper, labelled A-G, is provided along with the criteria each *Related Work* covers. Papers used are listed in Appendix A, and copies of the materials used are available on request from the author of the thesis.

- **A**: This paper was about a new machine learning approach to identify and resolve Chinese zero pronouns. It describes a limited number of relevant previous works, pointing out limitations in some of these. There is no

relation of cited work to the author's work.

- **B**: This paper is looking for culturally shared common beliefs and compares Chinese and English similes as a way to identify stereo-typical descriptions that exist between the two cultures. It is very descriptive with a variety of cited works and their limitations, although it has minimal reference to the author's work. The main issue is that most of the citations are not relevant given the *Abstract* or *Introduction*.
- **C**: This paper studies the extraction of entailed semantic relations through syntax based comma resolution. It is well written with descriptions of cited work highlighting gaps and how cited works relate to the author's work.
- **D**: This paper applies machine learning to discover semantically related pairs of words by using dependency relations. There is very little cited work and a lack of depth to the description and evaluations. The author's work is compared to only one cited work.
- **E**: This paper is concerned with automatic image annotation using auxiliary information found through image capture on the web using captions and keywords. It is well written describing the context, cited works and a very clear paragraph on how their work differs.
- **F**: This paper presents an annotation free approach to detecting foreign inclusions when parsing German. Although well written, this paper has a limited number of citations and never mentions the author's work, failing to put any of the cited works in context or show the author's novel approach.
- **G**: This paper uses an axiomatic approach to exploit lexical resources for query expansion to improve retrieval performance. It provides a limited amount of cited work and background along with one sentence on the author's work but limited relation between this and previous work.

The modified *Related Works* were reviewed and criteria there were assessed on included:

- **Style**: assessing the grammar and whether the text flows, e.g. good use of connectives compared to a bullet list style of citations.
- **Thoroughness**: are the citations appropriate given the introduction; are there enough citations; is there enough discussion on the cited work.

	Style		Thoroughness			Context		Crit Eval
	Grammar	Flow	Relevant	Enough	Discuss	Compare	Contribution	Limits/Merits
A	✓	X	✓	✓	X	X	X	✓
B	✓	✓	X	✓	✓	X	X	✓
C	✓	✓	✓	✓	✓	✓	✓	✓
D	✓	X	✓	X	X	X	X	X
E	✓	✓	✓	✓	✓	✓	✓	✓
F	✓	✓	✓	X	✓	X	X	✓
G	✓	✓	✓	✓	✓	X	X	X

Table 3.1: Expert assessment of present and missing criteria in the *Related Work*. Style (grammar, flow), Thoroughness (relevant citations, enough citations, enough discussion), Context (relations cited works to author and author contribution), Cited Evaluation(limits and merits)

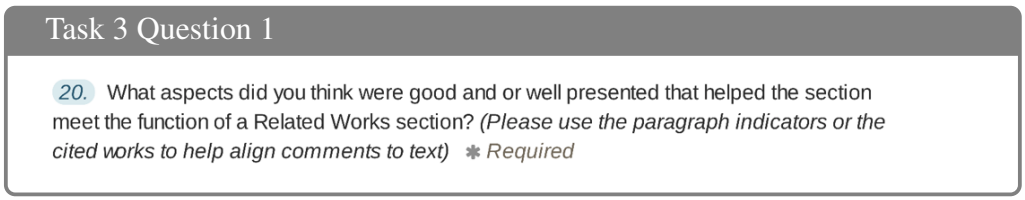
- **Context**: is there comparison of cited works to the author's work; is the author's contribution and the gap they are filling clear.
- **Critical evaluation**: do they highlight any merits or limitations for the cited works.

Table 3.1 shows the assessment of Related Works A–G with respect to these criteria, where X represents that the criteria was missing.

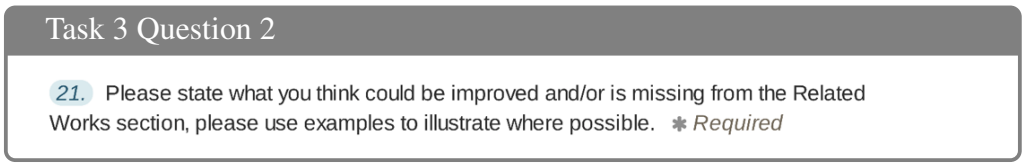
3.3.4.2 Peer Review of *Related Works*

Following the completion of the pre-task opinion questionnaire, the participant was taken to the main task. Participants were asked to read the *Title*, *Abstract* and *Introduction* of the *Related Work*. Then they were asked to describe in text or bullet points what they expected the *Related Work* to cover. The point of this was to focus the participant's mind on what they had read, but we do not use the responses during the analysis. Then followed the main questions for the task. Participants were asked for 2 free-text responses about aspects of the *Related Work*, and one multiple-choice question which included multiple ratings. We asked the open questions first in order to not to preempt the participants with the closed rating responses they may give. This was repeated for all seven examples of *Related Work* sections (labelled A–G, each comprising of *Title, Abstract, Introduction, and Related Work* section). The free-text questions asked:

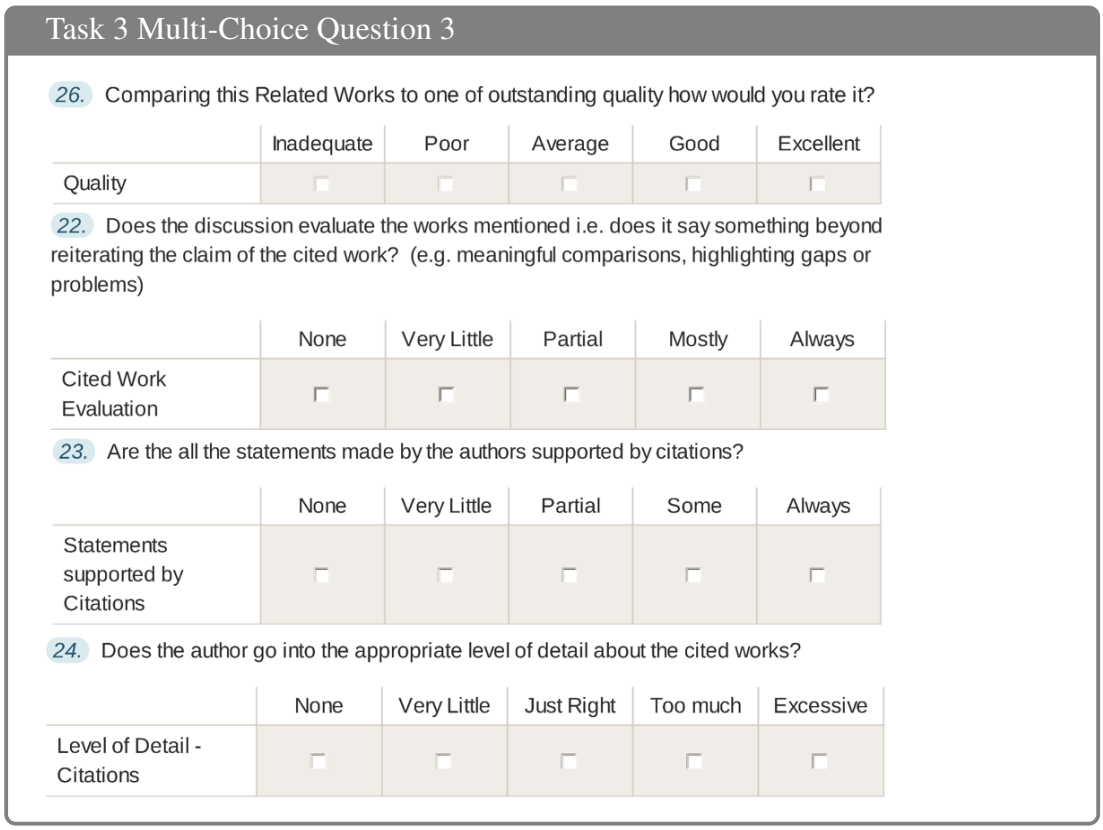
(1)-What was good or well presented



(2)-What could be improved and/or was missing in the *Related Work*.



(3) A multiple-choice question asked for ratings (on a 5 point Likert Scale) for each *Related Work* on four aspects: overall quality, context given in terms of meaningful comparison of author's work to cited works and evaluation, the detail provided for cited work, supported statements, i.e. were citations used to support statements where required. (Figure 3.3). As previously mentioned these questions were derived both from resources describing content that should be present in a *Related Work* and from using the rubric to measure PhD literature reviews in (Boote and Beile, 2016). Gogolin and Stumm (2014) mentioned earlier, which looks at developing a questionnaire for assessing the quality of articles during peer-review, highlight the importance of having scales greater than 4 to capture variation, hence our decision to use a 5 point Likert scale.



Task 3 Multi-Choice Question 3

26. Comparing this Related Works to one of outstanding quality how would you rate it?

	Inadequate	Poor	Average	Good	Excellent
Quality	☐	☐	☐	☐	☐

22. Does the discussion evaluate the works mentioned i.e. does it say something beyond reiterating the claim of the cited work? (e.g. meaningful comparisons, highlighting gaps or problems)

	None	Very Little	Partial	Mostly	Always
Cited Work Evaluation	☐	☐	☐	☐	☐

23. Are the all the statements made by the authors supported by citations?

	None	Very Little	Partial	Some	Always
Statements supported by Citations	☐	☐	☐	☐	☐

24. Does the author go into the appropriate level of detail about the cited works?

	None	Very Little	Just Right	Too much	Excessive
Level of Detail - Citations	☐	☐	☐	☐	☐

Figure 3.3: The rating questions for each *Related Work*.

3.4 Participant Demographics

Twenty-three participants completed the study and were divided into two groups: experienced researchers and students. They all met the criteria for experienced participants, to have completed their PhD at least five years earlier and first-authored over ten papers. Student participants were all PhD students or first-year post-docs. There were eleven participants in the experienced group (five males and six female) and twelve in the student group (six males and six female). Experienced participants were spread across age categories from 30 to 60+. Students were mostly between 25-29 years of age, with two between 30-35, and one over 45 and a mean of four for published papers. The fields that the experienced participants had published in can be seen in Figure 3.4 and the student participants in Figure 3.5. Experienced participants all had a minimum of 2 years of supervisory experience, and all had over five years of experience in reviewing articles. Only two students had supervisory experience and for reviewing articles: four participants had no experience, two less than one year, four participants between one and three years, two participants between three and five years. Six experienced participants were non-native English speakers, and four students were non-native English speakers. Of the experienced participants, only one had received no education (secondary, degree, MSc or PhD) in English and that same participant had not worked in an English speaking country. For student participants, only one had received no education in English (secondary, degree, MSc or PhD) and had worked for less than a year in an English speaking country.

3.5 Evaluation Methods

In this section we describe our approach to analysing the responses from our study, the methodological approaches taken and any statistical significance testing undertaken.

3.5.1 Peer-review Free-text Response

Taking a data-driven approach, Thematic Analysis is used (Braun and Clarke, 2006) to identify characteristics of the free-text responses. This approach means

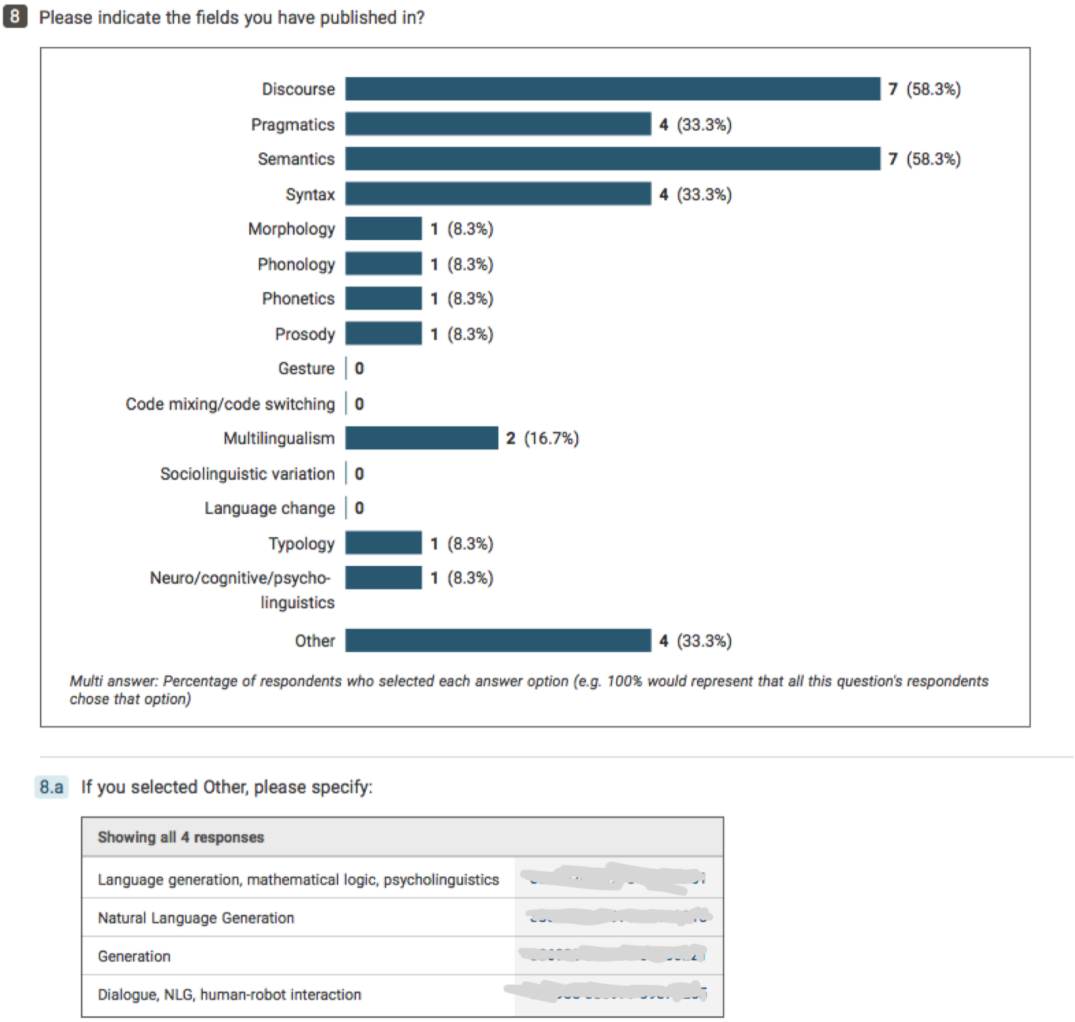
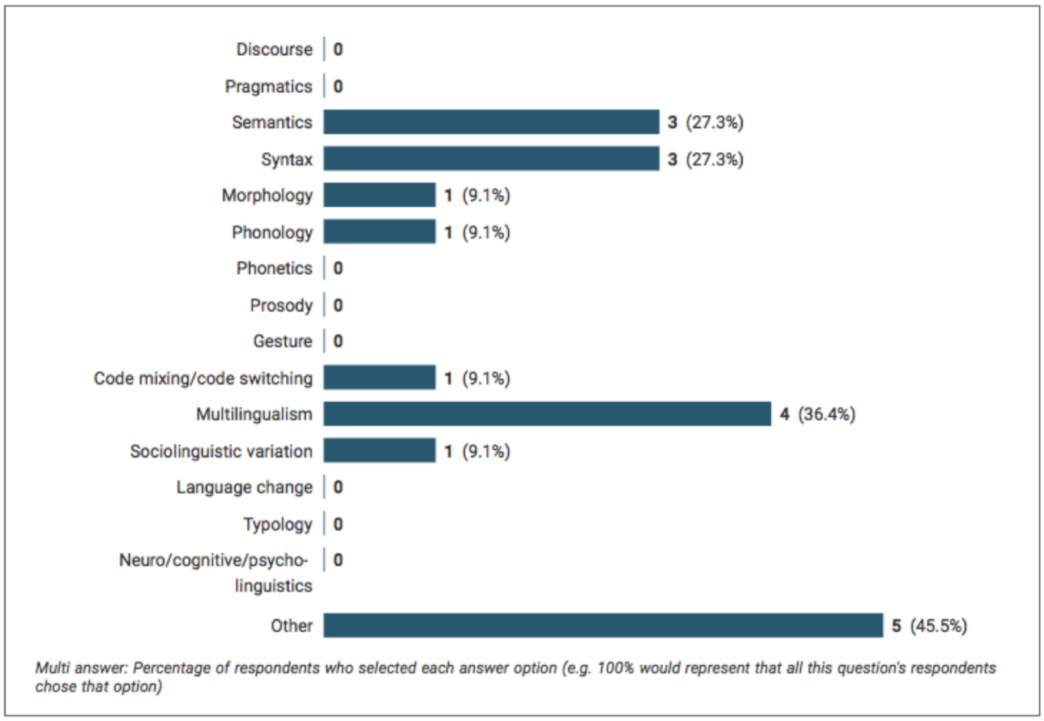


Figure 3.4: Fields from ACL paper submission listing, that the experienced participants had published in.



8.a If you selected Other, please specify:

Showing all 5 responses	
Music parsing, music information retrieval	
Language generation	
Formal language theory	
None	
Machine Translation	

Figure 3.5: Fields from ACL paper submission listing, that the student participants had published in.

themes emerge from the data (Patton, 1990), allowing coding of data whilst trying to avoid pre-conceived ideas or researcher bias (Braun and Clarke, 2006). Through successive rounds, codes are developed and applied to the data, and codes are interpreted to develop themes that capture the important characteristics in the feedback text. It should be noted, originally we considered undertaking an NLP approach to understanding the aspects participants raised, but it did not provide useful results. This is most likely due to the small number of participants and *Related Works* used.

For further insight into how the students and expert group differs, Epistemic Network Analysis (ENA) (Shaffer et al., 2009) is used. ENA is a graph-based analysis technique that can be used to examine and model the cognitive connections made in discourse. Coded co-occurrences in the discourse are used to create a high-dimensional representation which is then projected to a lower-representation space through single-value decomposition. ENA allows comparison of the coded free-text responses and understanding of the connections made in the discourse through a visual network. By subtracting the two group networks, one can compare and visualise what each group focused on more.

ENA has previously been used in the educational domain to study student cognitive connections during problem-solving (Nash and Shaffer, 2011) and in studies of interactions in discourse produced in online discussions (Shaffer et al., 2016).

3.5.2 Rating Agreements

Collecting and assessing agreement is a time-consuming task, particularly if expert annotations are needed. Recent crowdsourced platforms, such as Amazon Mechanical Turk[1] provide an avenue to collect significantly more annotated ratings in a cheaper and faster way. Recent work has considered the impact of using annotations from non-experts and shows that methods can be developed compensating for the noise generated in non-expert labels (Dgani et al., 2018; Yang et al., 2019; Snow et al., 2008) and that sufficient levels of the annotated corpus by non-experts compensates for noise (Kwitt et al., 2014). In this work, however, we are explicitly interested in the differences between the experts and the non-experts, i.e. the PG student group. Agreement during peer-review rating is known to be

[1] https://www.mturk.com

problematic, with reviewers exhibiting different opinions (Lawrence and Cortes, 2014). Therefore, given the smaller numbers of participants, we could expect that our experts may not be in high agreement. Nonetheless, it is useful to understand how much in agreement they are and if this agreement could allow us to draw conclusions from our experiments.

Judgements of agreement are subject to bias based on participants perceptions, understanding of rating scales and in this case, experience and familiarity with the task of peer-review. Additionally, different perceptions of the distance between items on a rating scale can cause varying degrees of disagreement (Gwet, 2014). In order to show that the conclusions we draw from our study have validity and reproducibility, it is necessary to show that there is some level of agreement within our rating schemas. Agreement statistics used in Computational Linguistics are comprehensibly discussed in (Artstein and Poesio, 2008) where they show that in order to be able to compare studies, any agreement must be corrected for chance. This is the agreement achieved beyond that which one would expect by chance alone. Overall agreement in ratings within Computational Linguistic tasks, such as ours, are usually assessed using Fleiss's weighted kappa (Fleiss, 1971) to calculate the agreement among the participants. This statistic allows for multiple raters and multiple ratings. Additionally, using the weighted kappa allows for specifying that items in a rating scale such as 1, and 2 are more in agreement than 1 or 5. We use a linear weighting with Fleiss's Kappa.

Given the small sample size and non-normal distribution, Mann-Whitney U tests are used, and Medians with inter-quartile range (IQR) reported to compare ratings between the two groups. We compare the groups at three levels: per-document, across documents and ratings in general. Statistical significance mainly depends on sample size, and we have a small number of subjects and *Related Works*. It is therefore important to report the effect size of any significant result which will help the reader judge if this difference is meaningful (Mangiafico, 2019). We report effect with *Vargha and Delaney's A* (Vargha and Delaney, 2000).

3.6 Results

3.6.1 Opinion Questionnaire about *Related Work*

Q 1&2 Function and characteristics of the *Related Work* Section:

Based on the free-text responses to the first two questions in this section, three main themes were highlighted by all 23 participants:

1. Putting the author's work in context to the existing body of work
2. Showing work that is related, and
3. Highlighting the difference or contribution of the author's work compared to related work.

Six of the student participants stated that part of the function was to allow the reader to follow the paper or to give them a background for understanding it. None of the experienced group suggested this.

Q3 Rejection of a paper based on inadequate or missing *Related Work* Section:

Responses were split between both groups. One experienced participant said this was *Highly likely* with 5 saying it was *Likely* and the remaining 6 saying it was *Unlikely*. In the student group 6 said it was *Likely*, 5 *Unlikely* and 1 *Very Unlikely*

Q 4&5 Ratings, Importance of Characteristics in *Related Works*:

Current Citations, Thoroughness, and Context were all thought to be *Important* or *Very Important* by 19 out of the 23 participants, with the others rating these of *Average importance*. Opinions differ about *Critical Evaluation*, with 10 experienced participants rating this as *Important* or *Very Important*, while 4 students

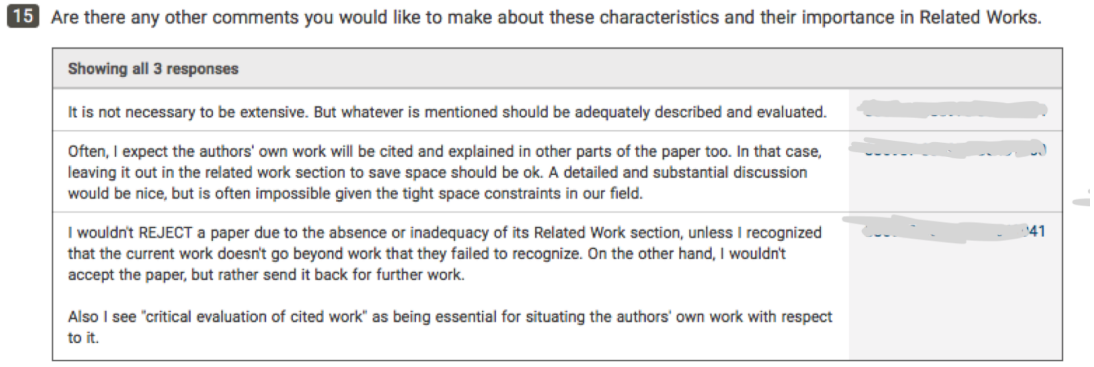


Figure 3.6: Responses from experts on comments on the characteristics of *Related Works* they were asked to rate.

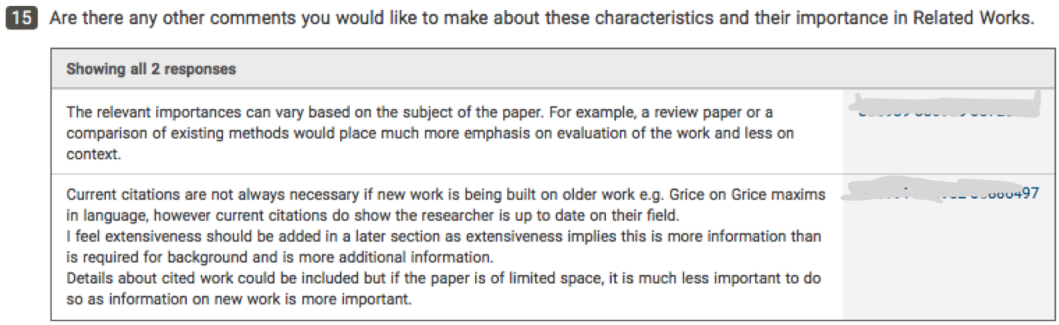


Figure 3.7: Responses from students on comments on the characteristics of *Related Works* they were asked to rate.

rated it of *Little Importance* and 7, of *Average Importance*. This indicates a difference in the students' view of citation evaluation. There is more evidence for this in the free-text responses discussed in the next section.

The question asking for any comments on these characteristics resulted in only three responses from experts and two responses from students (Figures 3.6 and 3.7). Experts commented on space being a problem for extensiveness of the discussion and the need for critical evaluation. One student commented on the subject of a paper influencing expected content within the *Related Work*. The other student seemed to think current citations were not necessary if the work was being built on older work, but information on new work was more important than detail of cited work if there was limited space.

Q6 Standard of *Related Works* in the last decade:

All student participants picked *other* for this question elaborating that they did not feel they had enough experience to comment. One experienced participant thought the standard had got better giving a reason that graduate students get more detailed advice. Three experienced participants though the standards had got worse. Their reasoning on why standards had declined centred on lack of space and authors only citing recent works. Three experienced participants thought standards remained the same and they offered no elaboration on why they thought this. These 3 participants were in the lower age category for experienced participants 30-40. Finally, four experienced participants chose *other* for this question. Comments focused on saying there are too many papers to read through and not enough space to adequately cite all work and the pace of the field (Computational Linguistics) is such that it generates so much new work there is no space to cite old work. There were other comments focused on reviewers being busy and what authors focus on due to the problem of publish and perish, such as:

> *"reviewers are very busy and especially the more experienced ones. It is time-consuming to make thorough reviews when you are given 4 or 5 papers for each A-ranked conference."*

> *"... the amount of time the authors can spend on related works. There is much more papers submitted to top-tier conferences because of the "publish or perish" mantra. More and more papers (most ?) belong to the salami-slicing category. The most valuable thing to do (in terms of acceptance rates) is the evaluation, related works are just becoming a secondary thing."*

3.6.2 Peer-Review - Free-Text Responses

Reviewing the qualitative responses in successive rounds of analysis led to the identification of 12 codes and these were categorised into four themes that reoccur in the responses, discussed below. Code names are the abbreviated terms given in bold below, e.g. **BG-Cxt, M-Com**.

3.6.2.1 Theme 1 – Context:

There were three areas where both groups look for context.
(BG-Cxt) - providing the appropriate amount of background context to situate the work in the field. Comments included:

"overview of knowledge acquisition was very helpful and added context"[s8]

"no hint of where this paper's approach is situated"[s11]

(M-Com) - involves meaningful comparison, which is the expectation that citations listed should be compared meaningfully to the author's work. Comments included:

"good comparison, comparing existing methods to current work is included"[s19],

"no attempt to relate the previous approaches to the current research"[s12]

(A-Cxt) - an expectation of acknowledgement of the author's problem through mentioning their contribution or the gap they are filling in relation to the cited work.

"clearly stated the novel contribution of this work"[s2],

"discuss how the paper's contribution differs from prior work"[s1]

3.6.2.2 Theme 2- Citations:

(C-Miss) - both groups highlight if they think citations are missing either from specific areas or in general if they do not think enough citations are present. Comments included:

"Overall this seems short and is missing citations."[s11]

(C-NRel) - An area the groups differed in is experts highlight when citations are not relevant to the author's topic, which is less likely to be highlighted by a student.
(C-Use) - both participants also highlight when citations are missing and needed

to show what the author has used or based their own work on. Comments included:

"more thorough description of the actual methods used"[s5].

(C-Eval) - in discussing the evaluation of citations, students look for an author to mention limitations or merits, particularly concerning technical details of the performance. Experts are less likely to mention this and focus more on suggesting evaluation needs to be shown to be relevant, i.e. put in context (coded above in context theme – **M-Com**). Examples of comments from experts:

"essential to investigate these to explain the novelty of this paper"[s19]

"why is it not sufficient to use these"[s9]

Examples of differing student comments:

"no details about limitations"[s4]

"Evaluation of references very basic"[s13]

"list of methods, no discussion of the merits and drawbacks of each"[s2]

3.6.2.3 Theme 3 - Discussion:

Both groups mention when there is:
(D-Max) - too much detail
(D-Min) - too little detail
(D-Right) - the discussion is thorough or comprehensive

Comments included:

"It seems to be the most related previous work, and it should be more thoroughly discussed"[s2].

3.6.2.4 Theme 4 - Language and Structure:

Both groups highlight when:
(L-Clear) - the language is not clear
(L-Struc) restructuring may be needed, but these are more frequent in student feedback. Comments included:

"I think the structure of this has problems with coherence ....using more simple sentences which are easier to understand would be helpful"[s16]

3.6.2.5 ENA Analysis of Free-text Responses

Epistemic Network Analysis (ENA) (Shaffer et al., 2009) is applied to consider the cognitive structures within the discourse, comparing the two groups based on the two free-text response questions - *what is good or well presented* and *what is missing or could be better*. The ENA networks are presented in Figures 3.8 and 3.9, codes on the graphs are the abbreviated codes from the four themes in previous section, e.g **BG-Cxt, M-Com**. The figures represent the subtracted network to show the connections each group focused on more and single points represent each participant, red experts, blue students. Experts focus more in *what was good* on highlighting meaningful comparisons between the author's work and cited works, along with the thoroughness of the discussion. The student group, on the other hand, focus more on saying the work is well situated within the field, i.e. background context is given, and that citations are evaluated. Similar to the expert group, students highlight the thoroughness of discussion and comparison. Comparing *what could be better or is missing* (Figure 3.9) experts' focus is on highlighting meaningful comparison of cited works to the author's work, that the author has demonstrated the gap they fill or contribution they make and on missing cited works. Students are focused on citation evaluation being missed, problems with the structure and insufficient detail.

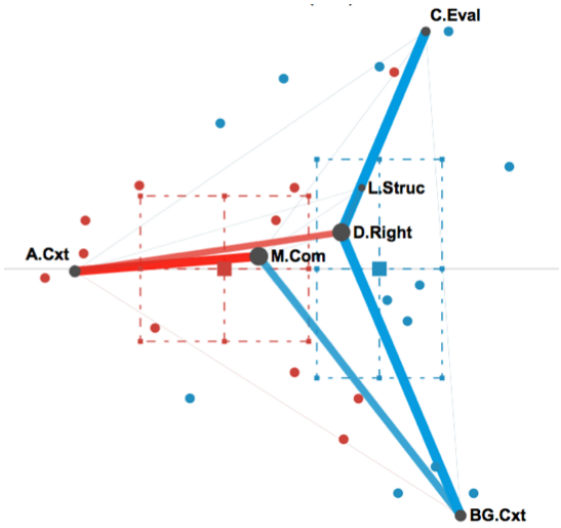


Figure 3.8: ENA free-text responses to what was good. Individual points represent each participant. The network is a subtracted network showing the connections each group focused on more, where Red indicates Experts and Blue indicates Students.

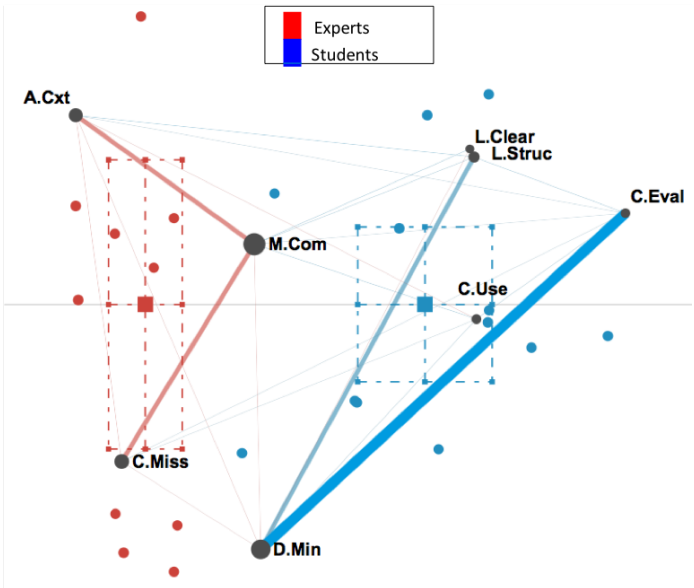


Figure 3.9: ENA free-text responses to what could be better or was missing. Individual points represent each participant. The network is a subtracted network showing the connections each group focused on more where Red indicates Experts and Blue indicates Students.

3.6.3 Peer-Review - Ratings

3.6.3.1 Agreement in Rating Between the Groups

The kappa statistic puts the measure of agreement on a scale where 1 represents perfect agreement, and 0 indicates agreement being no better than chance. Table 3.2 below shows the interpretation of kappa values and strength of agreement (Landis and Koch, 1977).

Interpretation of Kappa Value	kappa value
Poor agreement	<0
Slight agreement	0–0.2
Fair agreement	0.2–0.4
Moderate agreement	0.4–0.6
Substantial agreement	0.6–0.8
Almost perfect agreement	0.8–1.0

Table 3.2: Agreement interpretation for Kappa values (Landis and Koch, 1977)

Agreement	Quality	Context	Cit Eval	Support	Detail
Expert	0.40	0.45	0.35	0.36	0.41
Student	0.26	0.24	0.32	0.35	0.24

Table 3.3: Agreement (inter-rater reliability) for all *Related Works* by groups with Fleiss Weighted Kappa Fleiss (1971)

Agreement of the experts and students measured by Fliess's weighted Kappa (linear weighting) (Fleiss, 1971) can be seen in Table 3.3. The agreement is higher for all criteria in the expert group than the student group. Experts reach moderate agreement (as per Table 3.2) for Quality, Context and Detail, falling to fair agreement for Citation Evaluation and Support with students having fair agreement for all the criteria. Gogolin and Stumm (2014) modified criteria used for ratings during study iterations to match the criteria highlighted in open question responses, showing that agreement ratings then became more aligned. The need to align the rating criteria to the aspects experts specifically used to judge the *Related Work* could account for some of the disagreement between experts.

ID	Group	Quality	Context	Cit Eval	Support	Detail
A	Expert	2 (0)	1 (1)	2 (1)	4 (2)	2 (0)
	Student	2.5 (1.25)	2 (2)	3 (1)	4.5 (1)	2 (1)
B	Expert	2 (1.5)	2 (1)	3 (1.5)	4 (2)	3 (1)
	Student	**3.5*** (1.25)	3 (1.25)	4 (0.5)	5 (1)	3 (0)
C	Expert	4 (1.5)	4 (1.5)	4 (2)	4 (1)	3 (0.5)
	Student	4 (1.25)	4 (1)	5 (0.25)	5 (1)	3 (0)
D	Expert	2 (1)	2 (2)	3 (2.5)	4 (2)	2 (1)
	Student	2 (1.25)	3 (1.25)	4 (1)	4 (0.25)	2 (1)
E	Expert	4 (2)	4 (0.5)	4 (0.5)	4 (1)	3 (0)
	Student	4 (1)	4 (1.25)	4 (2)	4 (1)	3 (0)
F	Expert	2 (1)	1 (1)	3 (1.5)	4 (1.5)	3 (0.5)
	Student	**4*** (1.5)	**2*** (1.5)	4.5(2.25)	5 (1)	3 (0)
G	Expert	3 (1)	3 (2)	3 (1.5)	4 (2)	3(1)
	Student	3 (0)	3 (1.25)	3 (1)	4 (1)	3 (1)

Table 3.4: Agreement on ratings for all *Related Works* by groups. Medians are reported with Inter-Quartile Range reported in brackets e.g Median (Inter-Quartile Range), significance($p < 0.05$, Mann-Whitney U test) between groups denoted by *

However, we also know from our responses in this study students seem to struggle more from a pedagogical point of view about citation evaluation and understanding its value and function. Perhaps this lower agreement in the expert ratings is also related to a difference in understanding. We discuss more in the limitations Section 3.8.1 the value of further iterations to develop the rating criterion.

3.6.3.2 Agreement on Likert Ratings

Figure 3.10 gives an example of how students tend to rate higher than experts in overall quality, which was also true for context scores. Testing tendencies for ratings between the groups reveals quality, context and support are significantly higher for students, $p<0.05$ with Mann-Whitney U test. (**Quality** - U=2422, p-value=0.0023, **Context** - U=2581, p-value=0.0119, **Support** - U=2756, p-

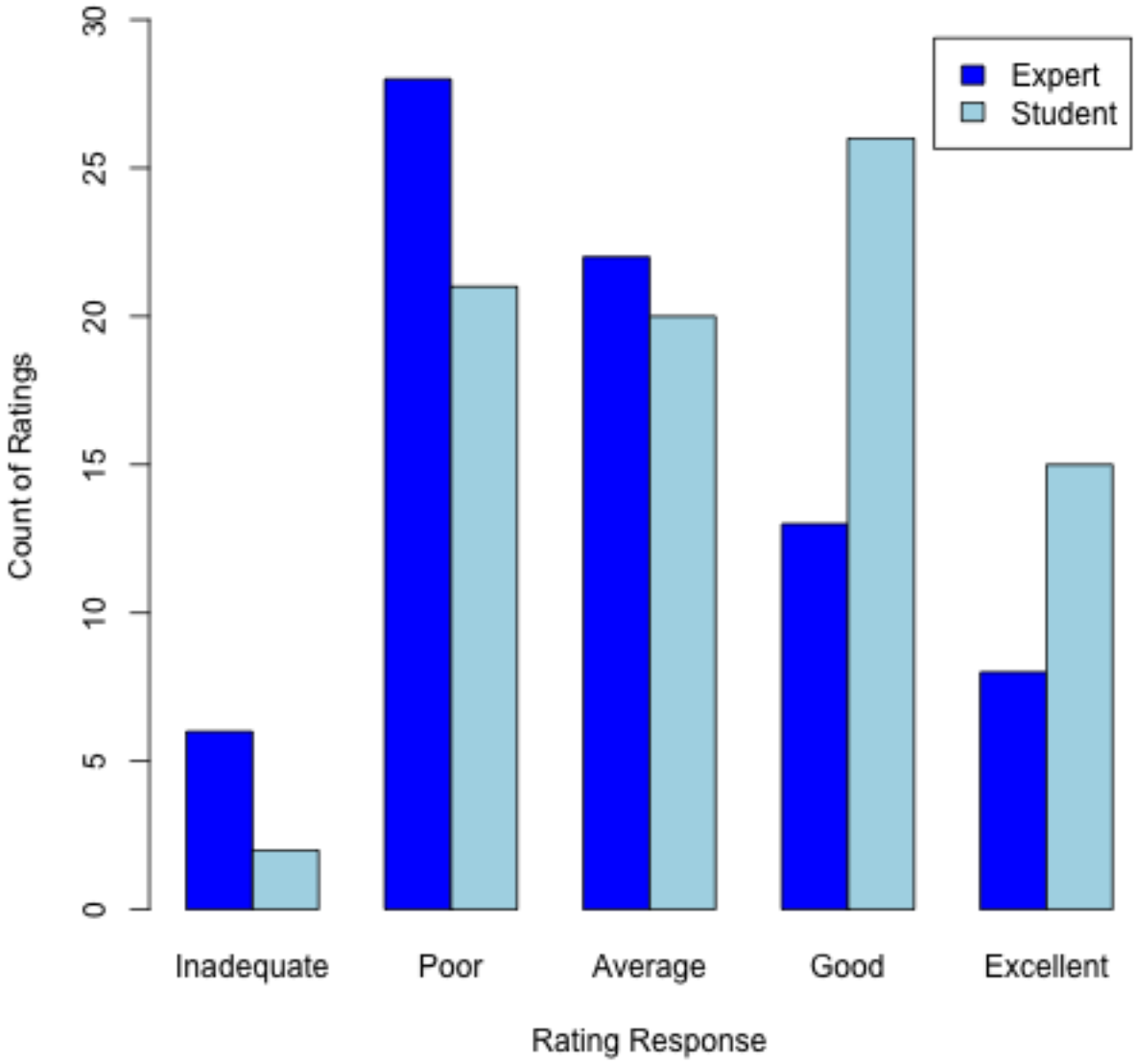


Figure 3.10: Counts for each of the 5 quality ratings (Inadequate, Poor, Average, Good, Excellent) by Expert and Student Groups, showing students tend to rate higher than experts.

value=0.0415). There were no significant differences between the groups in rating tendencies for citation evaluation and detail.

Ratings across all documents for averages scores on each criterion was compared using Mann Whitney-U and there were no significant differences. This is not surprising as from Table 3.4 we see that students and experts are in good agreement across Papers C, E, G. Disagreement seems to happen more often in the papers that have been described as having more features missing, (cf. Section 3.3)

Median ratings, Table 3.4, show several *Related Work* medians differ between students and experts, particularly for *context* and *citation evaluation*. However, only Paper B and Paper F differ significantly in *quality* ratings, with students being higher than experts, $p < 0.05$ with Mann-Whitney U test. (B - U=27.5, p-value=0.008) (F - U=35 p-value=0.024). Only Paper F differs significantly in *context* with students being higher than experts, $p < 0.05$,Mann-Whitney U test (U=35, p-value=0.023). In addition we calculate the effect size for each of these with *Vargha and Delaney's A*. Interpretation of *Vargha and Delaney's A* is described in Table 3.5, Quality rating difference Paper B, VDA = 0.82, large effect. Quality rating difference Paper F, between Experts and Students, VDA 0.90, large effect and finally context rating difference Paper F between Experts and Students, VDA= 0.60, small effect.

Interpretation of VDA Value	VDA value
Large	0.00 - 0.29
Medium	$>$ 0.29 – 0.44
Small	$>$ 0.44 – $<$0.56
Medium	0.56- $<$0.71
Large	0.71 - 1.00

Table 3.5: Agreement interpretation for Vargha and Delaney's A (Mangiafico, 2019)

3.7 Discussion

3.7.1 What are the content expectations highlighted in a *Related Work* by experts and do experts agree with each other?

The thematic analysis provides evidence that experts are consistent in the patterns of content they highlight as present or missing. Despite known problems of experts agreeing during peer-review, good agreement overall in quality ratings of experts is observed and good to moderate agreement in other ratings. Less agreement is seen in context and citation evaluation, but as discussed in the next section, both of these are aspects that students struggle with. It is possible that experts continue to struggle to recognise these aspects even after they gain experience, accounting for less agreement between the group. The agreements of what content experts expect to see in a *Related Work* can be summarised into four areas. The first being **Background Context** where an author is expected to demonstrate where their paper is situated within a field giving an overall context. Second, **Cited Works and Context** is expected to be given. The expectation is that there must be enough cited works, these should be relevant, and they need to be compared meaningfully to the author's work to demonstrate why they are relevant. Third, **Author Contribution** an author is expected to clearly identify their contribution by stating how their work differs from previous work. Finally, **Presentation and Language** there is an expectation that the *Related Work* should be presented clearly and be well structured. Whilst the aspect of discussion was highlighted - both groups mention when there is too much/little or comprehensive discussion - if the first three areas are covered adequately, the issue of providing enough discussion is addressed.

3.7.2 Do PG students differ from experts in what they look for in a *Related Work*?

In addition to the known problems of agreement during peer-review, students will have varied experience and are at different stages of development. These two aspects likely provide for more variation in student responses. Nonetheless, evidence in both free-text responses and ratings that students miss or differ in recognising characteristics when compared to the expert group is found. Overall

there was more agreement in the ratings for *Related Works* C, E and G between the groups. *Related Works* C and E were the better written and more comprehensive pieces covering what would be expected in a *Related Work*. *Related Work* G could be considered average, touching on most aspects but not covering all adequately. Students differed most to experts in context and citation evaluation and this likely impacted overall quality ratings. This is observed more in *Related Works* A, B, D, F.

Context: Students significantly differed from experts in their tendency to rate context higher. In discussing *what was good or presented well* or *what could be better and/or missing*, students put much less emphasis than experts on context in relation to the author's work with fewer comments on missing or comparison of the cited work to the authors **(M-Comp)** or that author contribution and context was missing **(A-Cxt)**. It appeared that students often did not recognise missing context or that the author's work had not been mentioned. This was particularly true in the case of *Related Work* F. F was written in an interesting and engaging style, and although like *Related Works* A and D it did not provide context between cited work and the authors work or author contribution, most students failed to notice this. This engaging style of writing in *Related Work* F seems to be more problematic for the students in recognising aspects.

Citation Evaluation: Opinion ratings and free-text responses suggest that students think about citation evaluation differently. Whilst differences in median ratings of citation evaluation between students and experts were observed, these were not significant. However, it does seem that students miss the deeper understanding of the purpose of citation evaluation that experts highlight, which is to provide evaluation with discussion on why this matters, i.e. what is the importance of this evaluation in relation to the field or the author's work. Students suggest evaluation only to list merits and drawbacks.

Quality: Students differed from experts by an overall tendency to rate quality significantly higher and in individual ratings for *Related Works* B and F. The difference of opinion about citation evaluation and ability to judge context likely contributes to the differences in making a judgement overall about quality.

Cited Work – Relevance: Unlike the expert group, students do not realise that the citations and most of the discussion in *Related Work* B are not relevant to the author's work, given the Introduction. This can be seen in free-text responses

and likely influences students' higher ratings in *Related Work* B.

Influence of Language and Style: There was a clear difference between the students and experts regarding free-text and ratings of *Related Work* F. *Related Work* F, like A and D, was a poorer *Related Work*, but it differed in that it was written in an engaging and interesting way, e.g. good transitioning between cited works with use of linkage words that provide signals to the reader (*however*, *on the other hand*, *conversely*). Previous work (Miltsakaki and Kukich, 2004) has shown that incoherence captured through rough shift pattern in entity transitions links to lower essay scores. *Related Works* A and D could be considered to have rougher shifts between cited works and students noticed more readily the issues with these works. Novick (1988) highlights that novices can get misled by superficial features, and it seems the more engaging style of writing does mislead the students, drawing their attention away from the characteristics they should be looking for during peer-review.

3.8 Summary and Limitation of the Study

The study undertaken gained an understanding of what content experts look for in a *Related Work* and if there is agreement between experts about these. It also gained an insight into where PG students differed from experts. Using this knowledge, the goal is to build a tool that can visualise and bring a student's attention to these content aspects. Whilst finding agreement between experts, the very thing that experts emphasise *context* is what the PG students appear to struggle with recognising, particularly in certain styles of writing. If they do not recognise these aspects are missing in reviewing others' work, it is likely they will miss this within their own writing. This does lead to the idea that there is perhaps more pedagogical intervention needed with PG students to teach about context and what it means in academic writing.

Whilst aspects of presentation, structure and language styles are highlighted, as mentioned in Section 2.4, these are important but are not part of the work undertaken in this thesis. However, the importance of these and how they may be incorporated in future work is discussed in the final chapter. Also not addressed in this work is if the citations are relevant given the *Introduction*, another aspect PG students appeared to struggle with.

Table 3.6 summarises the findings of expected content in a *Related Work* which is used as a framework in the next chapter in developing author intentions labels for *Related Work* feedback.

3.8.1 Limitations of Study

The work described in this chapter is subject to potential limitations, and the conclusions drawn should be considered in light of these limitations. Validity and reliability are important in showing that any study can be generalised to the population at large (Kelly, 2009, Ch. 12). Some aspects of our experiment design restrict our ability to be sure that our observations could generalise either to the Computational Linguistics domain or to *Related Work* section writing across all disciplines. The limitations centre mainly around our selection of participants and the small number of peer-review tasks. In selecting our subjects (cf. Section 3.3.1) we acknowledged that bias could be introduced by the subset of people asked, i.e. our participants were not necessarily reflective of the whole Computational Linguistics community. This can be seen in Figures 3.4 and 3.5, which show what sub-fields participants have submitted papers to. In addition, our sample size of seven papers was small and subjective, given that we manipulated the content, and this manipulation may be subject to our own biases.

Whilst these are potential limitations, we also argue that our findings are similar to what may have been expected from looking across relevant literature on how to write a *Related Work* section. Thus, we believe that, despite the limitations, there is value in the findings; particularly the findings where we were able to show how the PG students and experts differed. This understanding of where the PG students struggle allows for a better understanding of how pedagogy could be developed to help PG students in their *Related Work* writing.

There are several ways in which future work could extend or further validate the work done in this chapter. Firstly, a follow-up study with a larger and more diverse participant base would be valuable, and this should include both more experts and PG students. With a larger number of participants, an analysis could be done on demography aspects to understand if any of these, such as sub-field type, discipline or year of PhD affect responses. In order to reduce bias, an alternative approach to manipulating the *Related Work* section could be to collect a

corpus of draft and published material. Further analysis could also be done with respect to the rating criteria used in the peer-review task. The work of Gogolin and Stumm (2014) in developing criteria to judge quality took an iterative approach using feedback from early studies to improve and consolidate the criteria used to rate to a more comprehensive framework. In a future study, it could be valuable to consider modifying the rating criteria used in our study to align more to the findings of this chapter, i.e. what experts look for in a *Related Work*. In addition, the use of the criteria itself may bias what participants look for as they progress through the peer-review activities. It could be valuable to separate the tasks and compare results from groups using open and closed rating separately. This could allow for evaluating how the rating criteria influences the free-text responses.

Findings - Content experts expect to be present

Background Context the author is expected to situate their work in the field, demonstrating they understand their field and its history through indicating seminal works and relevant research fields.

Cited Works and Context There must be enough cited works and these must be relevant. This relevance is expected in terms of *Critical Evaluation* and/or *Meaningful Comparison*. Critical Evaluation makes clear the gaps/merits of a work and puts these in relation to the author's work. Meaningful Comparison shows how the cited works have influenced or are relevant to the author's. There are several ways this meaningful comparison can occur with the author explaining how:

1. the author's work differs in a specified way
2. the cited work is used or built upon by the author
3. the cited work is similar to the author's

Author Contribution Having exposed the gap, the author should identify their contribution or how their work differs.

Presentation through Language and Structure It is expected that the discussion must be structured and the language used should be clear. (These aspects are not addressed in this thesis.)

Findings - Aspects students struggle with recognising

Context PG students are less likely to notice, particularly if the *Related Work* is well written and engaging that the following are missing (1) the cited work is not made relevant to the author's work (2) that the author's contribution or how their work differs to any previous work (3) author's work is not mentioned at all. Whilst PG students understand that work cited should be critically evaluated they focus on listing these merits/limitations and not on why they are relevant to the author's work.

Table 3.6: Summary of findings from experts on content that should be present in a *Related Work* and where PG Students Struggle

Chapter 4

Mapping Content Expectation to Author Intentions

4.1 Introduction

This chapter builds a model of author intention for writing support in a *Related Work* and discusses how this relates to existing author intention models. This model is based on the framework of content experts expect to see in *Related Work*, and PG students struggle with, from the previous chapter. We describe the annotation study which investigates if the author intention labels can be annotated with reasonable human agreement. The annotation work presented in this chapter is published in (Casey et al., 2019b).

4.2 Author Intention Labelling and Annotation Unit

Previous works have successfully proposed author intentions models, but none focus specifically on giving feedback for *Related Work* sections. Some models capture parts but not all elements of intentions in the proposed framework in this thesis, such as those that consider citation function (Teufel et al., 2006a; Angrosh et al., 2012) or argument zones reflecting author intentions (Teufel, 1999; Teufel et al., 2009). These, however, are designed for different purposes, such as summarising or information extraction (e.g. gene relations, knowledge claims). Thus, they also have labels that are irrelevant to a *Related Work*, e.g.

Conclusion. The one aspect that these approaches have in common is the need for annotated data based on task-orientated annotation schemes. When we look closely at how labels from other models are applied during annotation for their specific task (cf. Section 4.5), we observe whilst their labels may look to be matching descriptors they do not match the intentions within our model. This results in us proposing our own task-orientated annotation schema.

Most approaches to segmenting within scientific articles are flat and label the discourse into functional regions (Webber et al., 2012), with most using the sentence as an annotation unit. However, some that consider citation context use partial or several sentences and some works, particularly those based on argumentation theory (Toulmin, 2003), use units of discourse based on such theories as Rhetorical Structural Theory (RST). Using a sentence as an annotation unit could introduce challenges – for example, a given sentence could potentially serve two functions that may be better captured at the clause level. Two functions occurring within a sentence could lead to ambiguity for annotators and impact the consistency of annotation. Ambiguity arises when there is intrinsic difficulty in choosing the correct annotation (Versley, 2008). Annotator difference due these ambiguities does not necessarily mean that one annotator is wrong just that the annotation label could be interpreted multiple ways. Annotation of intention labels in this work is done at the sentence level. However, we observe ambiguity in interpretation by the annotator, but also resulting from a lack of clarity by the writer. We discuss the impact of this choice further in Section 6.6 when considering the feedback given after our labels are used in *LitCrit*.

4.3 Mapping Expected Content to Author Intention Labels

The expected content framework developed in the previous chapter centres on qualities the experts look for in a *Related Work*. These qualities need to be mapped into author intention labels that can be used at a sentence level. From the previous chapter, there were four areas a *Related Work* was expected to cover: Background Context, Cited Works and Context, Author Contribution and Presentation through Language and Structure. The latter aspect is not considered for feedback in this thesis work. The first three areas are discussed next and

the sentence labels which form the author intention model, capturing content in a *Related Work*, are described along with example sentences for each sentence label.

4.3.1 Finding 1 - Background Context

This section describes how we map the findings of what experts expect to see regarding background in a *Related Work* (Table 4.1) into author intention labels.

Background Context the author is expected to situate their work in the field, demonstrating they understand their field and its history through indicating seminal works and relevant research fields.

Table 4.1: Findings from Chapter 3 on what experts look for in a *Related Work* with respect to background context.

In providing background context, an author may make general assertions or observations about the field and describe work in general terms, providing citations as evidence or not. The labels proposed capture evidence by noting when a citation is present. The reason for this distinction of whether the evidence in the form of citation is present is that novice writers are known to make limited use of citation types (Thompson and Tribble, 2001). In addition to a general assertion about the field, the author may also highlight a positive or strength/advantage in the field. The author may also highlight a gap in the form of a limitation or unaddressed area in the field. Evaluation of strengths and limitations help to distinguish between sentences that are descriptive only and those that are more informative. We create four background labels: two that capture background sentences with evidence, BG-EP and without evidence BG-NE; two labels that capture when something evaluative is said about the background, a positive BG(+), or a limitation/gap is mentioned, BG(-). The labels, descriptions and example sentences are found in Table 4.2.

Background Sentence Labels

Label	Label Description and Example
BG-EP	Background sentences with citations (evidence) *Most of the previous works conduct structure alignment with hierarchical structures, such as phrase structures(e.g.Kaji,Kida & Morimoto,1982), or dependency structures (e.g., Matsumoto et al. 1993;Grishma, 1994)*
BG-NE	Background sentence no evidence *In general, current approaches to NE identification usually contain two separate steps:word segmentation and NE identification.*
BG(+)	Background sentence highlighting a positive/strength/advantage *Recently, statistical NERs have achieved results comparable to hand coded systems.*
BG(-)	Background sentence highlighting a gap/limitation *Finally, the machine learning-based model has also been investigated and current models of this type are based on supervised approaches(Ittycheriah et al.,2001;Ng et al., 2001) that are heavily dependant on hand-tagged question-answer training pairs, which are not readily available.*

Table 4.2: Author intention sentence labels for background context findings in Chapter 3. There are four labels the first two are for a sentence with description only about the background/field with evidence BG-EP, or without evidence BG-NE. The second two labels are when critical evaluation on the background/field is offered. A positive observation BG(+) or a criticism or highlighting of a gap BG(-). Example sentences are provided for each label.

4.3.2 Finding 2 - Cited Works and Context

This section describes how we map the findings of what experts expect to see regarding cited works and context in a *Related Work* (Table 4.3) into author intention labels.

To provide informative feedback, there is a need to establish the relevance of a cited work to the author's work or if this cited work is descriptive only in nature. The first cited work label accounts for description only of a cited work **CW-DESC**. Two evaluative labels are captured for a citation sentence, the first label captures when merit about a cited work is highlighted **CW(+)** and the second when a gap is exposed by highlighting a limitation **CW(-)**. In the previous chapter, PG students struggled to identify context and there are several labels to capture context sentences. **A-CW** captures when a sentence directly compares a cited work and the author's work saying what is different. **A-SIM** captures when the author's work is similar to a cited work, **A-USE** for when a sentence that says the author uses/builds on or adapts/modifies a cited work. Teufel et al. (2006b) describes a category CoCoXY that contrasts two pieces of cited work and we capture this as **CW-COM**. The labels, their descriptions and example sentences can be found in Table 4.11.

Cited Works and Context There must be enough cited works and these must be relevant. This relevance is expected in terms of *Critical Evaluation* and/or *Meaningful Comparison*. Critical Evaluation makes clear the gaps/merits of a work and puts these in relation to the author's work. Meaningful Comparison shows how the cited works have influenced or are relevant to the author's. There are several ways this meaningful comparison can occur with the author explaining how:

1. the author's work differs in a specified way
2. the cited work is used or built upon by the author
3. the cited work is similar to the author's

Table 4.3: Findings from Chapter 3 on what experts look for in a *Related Work* with respect to cited work and its context to the author's work.

Cited Work and Context Sentence Labels

Label	Label Description and Example
CW-DESC	Cited Work description *Green, (2007) identifies argument structures in the biomedical field.*
CW(-)	Gap or limitation of the cited work is highlighted *However, they do not study the extraction of entailed relations as a function of the comma's interpretation.*
CW(+)	A positive/strength/advantage of the cited work is highlighted *Liu et al. (2004) used WordNet for both sense disambiguation and query expansion and achieved reasonable performance improvement.*
CW-COM	Two cited works are compared *Whereas Almuhareb and Poesio succeed in identifying the range of potential attributes and values that may be possessed by a particular concept, Veale and Hao succeed in identifying the generic properties of a concept as it is conceived in its stereotypical form.*
A-CW	Cited work and author's work are compared *In contrast to Kaisser(2006), we model the semantic role assignment and answer extraction tasks numerically, thereby alleviating the coverage problems encountered.*
A-USE	Author's work build on/adapts or uses the cited work *This method is also adopted in our system for non-peer phrase re-ordering.*
A-SIM	Sentence says that the author's work is similar to the cited work *Like our method, research which is based on the assumption of sentence alignments for parallel corpora has been done (Kaja and Aizono, 1996; Fung, 1997).*

Table 4.4: There are seven possible labels, CW-DESC when only explanation about a cited work occurs, positive evaluation CW(+) or a criticism/gap CW(-) about cited work. CW-COMP when two cited works are compared. A-CW when cited work and the authors work is compared. A-USE for the author's work builds/adapt, A-SIM author's work is similar to a cited work.

4.3.3 Finding 3 - Author's Work

This section describes how we map the findings of what experts expect to see regarding author's works in a *Related Work* (Table 4.5) into author intention labels.

Author Contribution Having exposed the gap, the author should identify their contribution or how their work differs.

Table 4.5: Findings from Chapter 3 on what experts look for in a *Related Work* with respect to the author's work.

These labels identify where the author(s) of the paper specifically mention their own work. There are three categories for sentences that discuss the author's work. **A-Desc** is a sentence where the author describes their work only. The label **A-Diff** was added after pilot annotations, as authors say *our work differs from previous work* with no explanation or linkage to what is different within the sentence. This is different from the label in the previous section (**A-CW**) where the author compares their own work and a specific work(s) to say what they do differently. Finally, **A-Gap** where the author highlights the novelty or points to a gap they fill in a sentence. A reader's experience may allow them to interpret when a discussion about the author's work is a description only sentence (**A-DESC**) rather than a sentence discussing their contribution (**A-GAP**), i.e. when it is not linguistically marked, such as when an author says explicitly *our contribution is* within a sentence. Distinguishing between these two categories of A-DESC and A-GAP does, however, prove challenging at the annotation stage and in the automated recognition of labels. Author labels, their descriptions and example sentences can be found in Table 4.6.

Author Sentence Labels

Label	Label Description and Example
A-DESC	Describes the author's work. *In our framework, we integrate Chinese word segmentation and NE identification into a unified framework using a class-based language model.*
A-GAP	Captures when an author specifically say their work is novel, new or describes how they address a gap. *However, since our method caught extracting the translation pairs as the approach of the statistical machine learning, it could be expected to improve performance be adding new features to the translation model.*
A-DIFF	Author's work is different (no information on how it differs). *Our work differs from previous work.*

Table 4.6: Intention sentence labels for author findings in Chapter 3. There are three labels, the first is for a sentence that describes the author's work A-DESC. The second is where an author mentions specifically the gap they fill or the novelty of their work A-GAP. The third label is for when an author says their work differs but does not provide an explanation as to how A-DIFF.

	A	B	C	D	F
1	DOI-l	SENT_ID	SENTENCE	SENTENCE_COREF	LABEL
525	D07-1061	417010	© 2007 Association for Computational Linguistics | many WordNet-based measures of lexical similarity based on paths in the hypernym taxonomy .	\xA9 2007 Association for Computational Linguistics 2 many WordNet-based measures of lexical similarity based on paths in the hypernym taxonomy .	OCR

Figure 4.1: Screenshot from annotation screen showing an OCR error

4.4 Learning from Pilot Annotations

A preliminary annotation study was conducted which highlighted a problem when considering author differences. There were occurrences of an author sentence which just indicated *our work is different*, giving no details of why or how. The annotators pointed out that these were not very informative sentences and quite different from when the author provides details of why their work is different. The extra label, A-DIFF, was added to account for this.

In addition, there were some sentences which had OCR problems, so a category was created for this, along with a category for TXT. The criteria for an OCR error was when the text in a sentence had become garbled and no longer made sense, such that a label could be applied. This happened usually when moving from the bottom of a column to the next column in a paper and a footer was placed in the text and some text missed, or when captions from a figure were interjected into the text. Figure 4.1 gives an example of an error that is created during the PDF to text process. In this example, a footer from the journal the paper appears in is written into the sentence text and some of the original sentence text is missing. TXT indicates that an author says *In the next section we will discuss*. This type of category was in the original AZ schema (Teufel, 1999), but it was thought it unlikely to arise in a *Related Work* section. However, it was highlighted in the pilot annotations. A category of OTHER was also added as there were some sentences the annotators could not assign a label.

4.5 Relating the Annotation Schema to Existing Works

Looking just at label names in the proposed schema, it would seem that these are direct replications of other models. However, on closer inspection of how authors apply these labels, discrepancies are often found that would not work

for *Related Work* feedback. One contributing factor as to why existing labels do not adequately support the goals in this work is that they are designed to look across the whole of a document. As a result, they seek either very general or much finer-grained labelling than required in this work. For example, Fisas et al. (2016) distinguishes between an author using data or using tools from another cited work. This finer-grained approach is not relevant or needed to provide feedback in a *Related Work* section.

A comparisons of the labels in our schema, which we call **LitCrit**, is carried out to those that are most closely related and were described in Section 2.2.1 - Argument Zoning described in (Teufel, 1999) and AZ-II (Teufel et al., 2009), CoreSC (Liakata et al., 2012), ArguminSci schema described in (Fisas et al., 2015, 2016) and citation function work of (Angrosh et al., 2012) and (Teufel et al., 2006a).

Three comparison tables are presented, Table 4.7 which compares background labels, Table 4.8 which compares the cited work labels and finally, Table 4.9 which compares the author labels.

LitCrit Label	Comparison
BG-NE BG-EP	All the intention models use a label of *Background* but they do not distinguish between those that have citation evidence or not. There are some discrepancies in what these capture to LitCrit, e.g. in Angrosh et al. (2012) this is used for *sentences that provide background or introduction*. Fisas et al. (2016) in addition to sentences that state common ground includes sentences of previous related work in their background category. The reason for their more general approach could be attributed to these other works capturing labels across the whole article.
BG(+)	We did not find evidence of other works looking for strengths in background sentences.
BG(-)	Teufel et al. (2009)work is the only evidence of where we can find a similarity to LitCrit's label of a shortcoming in the field although her label *GAP_WEAK - lack of solution in field, problem with other solutions* covers a shortcoming in both the field and a cited work.

Table 4.7: Comparison of background author intention labels from the schema in this thesis to other existing author or citation function models.

LitCrit Labels	Comparison
CW-DESC	Teufel et al. (2006b) and Fisas et al. (2016) have a category *Neutral* which is directly related to the LitCrit category of CW-DESC. These are used like CW-DESC label for descriptions of a cited work. Fisas et al. (2016) differs slightly in that they also include in this category *references for more information* or *comments on common practices* which we would put in one of the Background sentence labels. Teufel et al. (2006b) also allows this label to be used for an *unlisted citation function* or *not enough evidence to put in any other category*. In LitCrit these would go into the *OTHER* label. Angrosh et al. (2012) provides two labels *RWD_CS – a sentence describing a citation occurring in that sentence*, *RWD – a sentence describing a related work where the citation does not occur in that sentence*. LitCrit's one label covers both of these labels.
CW-COM	Teufel et al. (2006b) includes a category CoCoXY which contrasts two pieces of cited work as the LitCrit sentence label does.
CW(+)	Angrosh et al. (2012) has two labels that represent what LitCrit captures here RWS_CS and RWS. The first of these labels mentions a positive (strength) in a citation sentence and in the second a positive (strength) is mentioned but the citation is not present in that sentence. Fisas et al. (2016) also has this label CRITICISM-Strength.
CW(-)	The evaluation category for cited works relates directly to (Teufel et al., 2006b)'s category of *Weak - weakness of cited approach* and Fisas et al. (2016)'s Criticism-weakness. Angrosh et al. (2012) labels this as *RWSC - sentence noting the shortcomings in the related work citation.*

Table 4.8: Comparison of cited work intention labels from the schema in this thesis to other existing author or citation function models.

LitCrit Labels	Comparison
A-GAP	This has similarities to Fisas et al. (2016)'s Novelties, although their label is not exclusive to the author's approach and could include other cited work. Teufel et al. (2009) 's category of NOV-ADV is for sentences claiming a novelty or advantage of the author's own approach
A-CW	LitCrit category of author and cited work comparison, directly relates to the category of Fisas of Comparison-difference.
A-DESC	We could not find a schema that labels sentences just as author description. Other works such as Teufel et al. (2009) have several labels which in part fall under this category such as :OWN_MTHD, OWN_FAIL,OWN_RES,OWN_CONC, AIM. These are very specific and likely not to occur very often in a *Related Work*.
A-SIM	Both Fisas et al. (2016) with a label of *Comparison-similarity* and Teufel et al. (2006b) with a label of *PSim* have categories that label sentences with *authors work is similar to the cited work*.
A-USE	(Teufel et al., 2006b) and Teufel et al. (2009) have labels which align with this label of A-USE. However, they break this into finer detail than is felt necessary for the writing goal. Fisas et al. (2016) has four labels for using another cited work: *Use-method, Use-Data, Use-Tool, Use-other* and three labels for authors work based on a cited work, *Basis-previous own work, Basis Others work, Basis -future work*. Teufel et al. (2006b) has three labels: *PBas, uses cited work as basis*, *PUse, author uses tools/algorithms/data/definition*, *PModi, author adapts or modifies tools/algorithms/data*. This finer grained labels supports the goal of these authors as they look across a whole document but is not necessary for the goal of writer feedback.
TXT	In her original AZ schema Teufel (1999) includes a label of TEXT that is the same as the LitCrit label.

Table 4.9: Comparison of intention labels talking about the author's work from the schema in this thesis to other existing author intention or citation function models.

4.6 Corpus Description

Although at the annotation stage a label is assigned to a sentence, in subsequent chapters it will be necessary to look at all sentences related to a citation to determine what feedback to give and to produce better automated labelling. Understanding where co-references to citations occur in our data is, therefore, critical to providing feedback. The work in this thesis is not about solving co-referencing, thus, a data set that was already marked for co-reference to citations and author's work was chosen (Schäfer et al., 2012).

The corpus used for annotation is from (Schäfer et al., 2012) consisting of 266 published scientific papers from the ACL anthology (Bird et al., 2008). Their data set was extracted from PDF by commercial OCR software, sentence-tokenised and then manually annotated, using MMAX2 (Müller and Strube, 2006) for co-references. All the papers were 6 to 8 pages long. This is important as short-conference papers (4 pages) would have considerably shorter *Related Work* sections. We processed the full data set but only those papers with *Related Work* sections were extracted. This resulted in a data set of 113 papers. The final data set was comprised of the 95 *Related Work* sections that remained after papers with OCR problems were removed.

4.6.1 Extracting Co-references from the Data

The core annotation task carried out in (Schäfer et al., 2012) was to detect and track all mentions of an entity and put these into equivalence classes with all reference to the same entity tracked and linked. Each entity type, when mentioned, was put into one of 8 *Mention Classes* by an annotator. The class *ne* holds all proper names including citations and subsequent references to different *Mention Classes* were linked. Assuming that all cited work would be first mentioned as a citation, we use the class of *ne* to find all initial citations for each paper and all subsequent mentions across the mention classes are linked. Only entries for citations that exist in the *Related Work* section are extracted. A co-reference only exists for a citation if there are at least two mentions. Single citations are therefore not highlighted. A parser is built to identify any remaining single use citations, described in Section 5.3.4. Finally, references to the author's own work (within the paper not other work the author may have done) are also part

of the annotated data of (Schäfer et al., 2012). Using all the pronoun mention categories any reference to the author's work is considered, e.g. *our work, ourselves, we, our algorithm*. Looking for the first instance that occurs in the paper for these, the remaining references are extracted for the *Related Work* sections.

4.7 Annotation Study

This section describes the annotation study carried out with the author intention model.

4.7.1 Annotators

Both annotators were PhD students in Computational Linguistics, in the final stages of their degree programs. As knowledge possessed by researchers in a field can (in some instances) be used to overcome a lack of explicit linguistic marking, PhD students were preferable over domain experts in terms of bringing some, but not a lot of, knowledge to the task. This problem was acknowledged in Section 2.2.1.2 and highlighted in (Teufel et al., 2009) who instruct their annotators to only use rhetorical linguistic knowledge, but point out how difficult it is for domain experts not to use their knowledge when annotating.

One annotator annotated the whole corpus and the other just over half the corpus (i.e., 53 *Related Work*) sections.

4.7.2 Annotator Task

The *Related Work* sections were given to each annotator in an Excel file. Each row represented a sentence, with fields corresponding to document id, sentence id, the original sentence, and the sentence with citation and co-references marked. In the following field, the annotator entered a label from the pre-populated list provided. The final field was for comments, or for indicating any annotations they were not sure about. A screenshot of the Excel file used for annotation can be seen in Figure 4.2

	A	B	C	D	F	H
1	DOI-I	SENT_ID	SENTENCE	SENTENCE_COREF	LABEL	COMMENTS
7	C02-1010	403809	Unfortunately , automatic parse-to-parse matching has some weaknesses as described in Wu (2000) .	Unfortunately , automatic parse-to-parse matching has some weaknesses as described in CIT1 .	CW-SC	
8	C02-1010	403810	For example , grammar inconsistency exists across languages ; and it is hard to handle multiple alignment choices .	For example , grammar inconsistency exists across languages ; and it is hard to handle multiple alignment choices .	CW-SC	
9	C02-1010	403811	To deal with the difficulties in parse-to-parse matching , Wu (1997) utilizes inversion transduction grammar (ITG) for bilingual parsing .	To deal with the difficulties in parse-to-parse matching , CIT1 utilizes inversion transduction grammar (ITG) for bilingual parsing .	CW-P	
10	C02-1010	403812	Bilingual parsing approach looks upon the parsing and alignment as a single procedure which simultaneously encodes both the parsing and transferring information .	Bilingual parsing approach looks upon the parsing and alignment as a single procedure which simultaneously encodes both the parsing and transferring information .	CW-DESC	

Figure 4.2: Screenshot of the annotation screen in Excel showing cells for: document id, sentence id, sentence, sentence with placeholders for citation and co-reference, annotation label - drop-down and finally for comments.

4.7.3 Annotator Support

The annotators were given nine pages of guidelines (cf. Appendix C) which contained examples and suggested workflow to decide on an annotation label. Initially, the annotators met to discuss the guidelines and ensure their understanding. They trained on the same ten *Related Work* sections and compared their results discussing any differences.

4.8 Annotation Results

4.8.1 Corpus Analysis

The annotated corpus includes 95 *Related Work* sections and a total of 1,806 sentences. Double annotation was done for 53 *Related Works* and 955 sentences. The size of the data set is comparable to others who have studied annotation of scientific publications. Fisas et al. (2015) studied a corpus of 40 documents, Teufel et al. (2009) studied 90 papers, Feltrim et al. (2006) 52 abstracts, and Anthony and V. Lashkia (2003) 100 abstracts.

The results discussion focuses on the part of the corpus that double annotation was completed on to show the inter-annotator agreement and highlight the challenges. The annotated corpus is available on request from the thesis author.

4.8.2 Measuring Inter Annotator Agreement

Cohen's k (Cohen, 1960) is used to measure the annotator agreement, correcting for chance agreement. Cohen suggested the Kappa result be interpreted as follows: values 0 as indicating no agreement and 1–20 as none to slight, 21–40 as fair, 41– 60 as moderate, 61–80 as substantial, and 81–100 as almost perfect agreement. The formula is:

$$K = \frac{P_o - P_e}{1 - P_e}, \tag{4.1}$$

where P_o is observed and P_e is expected agreement. The range of Kappa can be between -1 and 1, where 0 means agreement is only expected by chance.

Kappa measures are widely used in annotation agreement in scientific publications in models that have been successful in automated classification based on their annotations (Teufel et al., 2009; Liakata et al., 2012; Fisas et al., 2016). In general, work on author intentions that uses Kappa agreement reports agreement in a range of 65-78% (Teufel et al., 2006a; Fisas et al., 2015; Teufel et al., 2009) with (Liakata et al., 2012) being much lower at 55.

Teufel et al. (2009) points out that Kappa treats agreement in rare categories as surprising and rewards these more than frequent categories. Although she sees this as an advantage because scientific publications often have these rare categories, others see this as misleading and criticise that chance-corrected measures do this when applied to unbalanced data-sets. Hence, others often report raw agreement (Kirschner et al., 2015). The data used here does have rare categories, thus, raw agreement in addition to the Kappa agreement is reported.

4.8.2.1 Inter-annotator Agreement

The inter-annotator agreement (IAA) was 77% (N = 955, n = 53, K = 2). Raw agreement was 80.10%. These results demonstrate substantial agreement and are comparable to similar studies mentioned earlier.

Out of the 955 sentences doubly annotated, the annotators agreed on 764. Based on the agreed sentences, the most frequent category was CW-DESC (32.50%), followed by the background categories BG-EP (12.20%) and BG-EP (10.90%). Following this were the author categories A-CW (9%), A-SIM (8.80%), A-DESC (5.80%) and A-GAP (3%). In the next section, some of the difficulties

the annotators had with A-CW versus A-GAP/A-DESC are discussed. CW(-) was surprisingly infrequent at 3.90% and CW-COMP at 2.23%. OCR and OTHER were both 1.30%. All the remaining categories constituted less than 1% of sentences, and interestingly all of these had good agreement - CW(+), BG(+), BG(-), A-USE, TXT, A-DIFF. OCR will not occur in writing feedback as text from PDF will not be processed. However, OTHER or TXT could happen, although these were rare categories with TXT having 13 sentences in agreement and OTHER 10 sentences in agreement. TXT was almost in perfect agreement, while OTHER was used more frequently by one annotator.

	A-CW	A-DESC	A-DIFF	A-GAP
A-CW	69	8	5	7
A-DESC	1	44	0	1
A-DIFF	-	-	2	-
A-GAP	5	6	2	23

Table 4.10: The agreement matrix between the annotators for author intention labels.

	BG-EP	BG-NE	CW-DESC
BG-EP	83	10	16
BG-NE	2	93	6
CW-DESC	6	5	248

Table 4.11: The agreement matrix for the annotators on cited work and background labels

4.8.3 Sources of Disagreement

There were two primary sources of disagreement between the annotators: one was in agreeing the labels about the author's work, and the other was in distinguishing between background sentences and those that pertained to specific citations.

In particular, the annotators noticed that when an author spoke about how their work was different to someone else's, they often broke this down over several sentences. The guidelines instructed the annotators to only mark what was linguistically indicated, but they were unsure if this meant in the text in general or in that particular sentence. This led to annotators disagreeing on A-CW and A-GAP/A-DESC, as can be seen in Table 4.10. Annotation guidelines need to be reviewed with some very specific examples that incorporate these scenarios with clear instructions on how to take linguistic markings into account. This will be a challenge for automated classification of the labels and in writing feedback. It needs to be considered carefully how this lexical information which occurs in previous sentences can be captured.

In disagreement about background sentences compared to citation sentences, seen in Table 4.11, one annotator highlighted that some sentences talked about two specific citations and they labelled these as BG-EP, while the other annotator labelled it as CW-DESC. After discussion, it was suggested that including examples of this kind in the annotation guidelines would have helped.

Annotators also noted that a sentence might belong to two labels. For example, a sentence may say something positive about a cited work but then highlight a shortcoming. In the guidelines, annotators were instructed to choose the author based labels over cited work labels and limitations which expose gaps over positives. In choosing the sentence as the annotating unit, it was acknowledged this could occur and is discussed more in our classifier error analysis Section 5.8.

There were two *Related Work* sections that included references to systems by their names, e.g. Moses or U-SVM. The annotators struggled with both of these as they were only given the *Related Work* section. If they had the full paper, they thought they would better ascertain if the author were referring to something that was their own work or another person's. One annotator questioned whether these types of *Related Work* were more likely to come at the end of a paper once a reader was familiar with these terms. Neither annotator thought the guidelines could be updated as in this instance, it would have been better to have access to the full paper. Again, this is going to be a challenging area for any automated system, especially if it only takes a submission of the *Related Work* section into account. The system will have no way of knowing if phrases of this kind relate to the author's work. It also raises a point that although this work is within one

discipline can sections still be written in different styles. Prior to this comment, it had not considered if order within a document impacted the style of the *Related Work*. However, it should still fulfil the qualities expected.

4.8.4 Annotating the Remaining Sentences

Following a discussion between the annotators on labels that were not in agreement, some changes were made. A small number of the disagreements were genuine mistakes with an annotator selecting the wrong label, but most were about the differences in A-CW versus A-GAP/A-DESC, and between CW-DESC and the Background categories. This resulted in an increase in Kappa agreement to 85% and raw agreement to 87.30%. One annotator carried out labelling of the remaining sentences following the discussion. The labels from the annotator who completed all sentences is used as the standard in the next chapter for automating label classification.

4.9 Summary

This chapter described how we built our model of author intentions, mapping these from the findings in Chapter 3 on what experts look for in *Related Work*. The labels were described in detail and compared to previous work showing differences in the annotation of what appear to be similar labels or showing where agreement existed in how labels were annotated. The corpus used for annotation was described, and good agreement was reached in our annotation study of 77%. Challenges in agreement exist though, and this is particularly true in establishing agreement between sentences that describe contributions and those that only describe the author's work. Also challenging are some differences between background sentences and sentences that describe more than one cited work. We observe our classifier in the next chapter also has similar problems in applying these labels. The annotated data in this chapter forms the training data in the next chapter, which focuses on automating the recognition of the author intention labels.

Chapter 5

Automating Recognition of Author Intention

5.1 Introduction

This chapter describes the approach to automate the recognition of author intentions within *Related Work*. We start with a discussion about the reasoning behind taking a feature-based approach with a supervised classifier. We give an overview of our feature-based approach that is used to learn the author intentions and how these are motivated or related to previous work is described. We then present our classifier model results which we follow with error analysis on mis-classification. Using the error analysis, we implement improvements to the classifier to increase its performance. Parts of the work presented in this chapter are published in Casey et al. (2019c).

5.2 Approach to Classifying Author Intentions

Our goal here is to explore if author intentions can be automatically recognised, but we are also interested in understanding how features contribute to classification outcome or errors. This leads to a compromise between state-of-the-art performance and being able to understand and explain relationships between features and errors. Gaining insight into how features influence the labelling may support improved feedback on the writing. In Section 2.7, we highlighted

that recent advances in NLP have advantages over traditional approaches, which use hand-crafted features by making use, for example, of contextualised embeddings (Peters et al., 2018) or pre-trained models (Devlin et al., 2018; Beltagy et al., 2019). Pre-trained models allow unsupervised training on large corpora, and fine-tuning can be done using a much smaller, labelled data-set for the specific task. There are some challenges with these approaches, such as the fact that it is not always possible to gain a true understanding of the role of features in detecting correct labels, automatically learned features could be due to idiosyncrasies of the data. Another challenge is that the available pre-trained models are trained on corpora that differ from ours, such as BERT which is pre-trained on English Wikipedia and BooksCorpus (Devlin et al., 2018). The work of Beltagy et al. (2019) show that there is only a 42% overlap on vocabulary between the scientific domain and these pre-trained models. Discussing previous work in Section 2.2.3 we saw that the use of pre-trained models on relevant corpora was what contributed to the strongest models. Obtaining corpora relevant to ours for training would be significantly time and resource-intensive.

Despite their wide adoption and performance in NLP tasks, neural approaches do not always provide the state-of-the-art. This has been observed in stance detection work. Stance detection is similar to parts of our task, in that for citations occurring in the *Related Work* we are detecting whether a stance is taken by labelling these sentences as evaluative. Siddiqua et al. (2018) show that detecting stance in Tweets, taking a feature-based approach based on POS tags with a SVM classifier, outperforms the state-of-the-art neural approaches. In the SemEval task of detecting stance in Tweets, Mohammad et al. (2016) observe that the state-of-the-art models, many with neural approaches, do not exceed the baseline SVM with n-grams. Aldayel and Magdy (2019) improve the state-of-the-art for this same task using more innovative features, but still with a SVM linear kernel classifier. In line with us, they also argue that taking this approach allows them to understand the role of features better than, for example, a neural approach. We also see evidence in other fields that neural approaches do not always generate the state-of-the-art, such as in Clinical NLP: a rule-based approach outperforms a neural approach when predicting named entities and relations (Gorinski et al., 2019).

In their review of existing writing evaluation systems, Hussein et al. (2019) point

out that no commercially available system for writing feedback use neural approaches. Additionally, we did not find neural approaches used in either of the academic systems we discuss in Section 2.3. Those models that did mix neural and hand-crafted features only focused on the scoring of essays. This approach using hand-crafted features in education is likely due to the need to explain relationships between features and the outcome of models. This is seen in other commercial NLP applications, e.g. **TheySay** (Moilanen and Pulman, 2016), a tool for sentiment analysis.

So, whilst neural approaches have made a significant impact on state-of-the-art, we choose to take a feature-engineered approach. Our main reason for doing this is to learn about the features themselves and how they contribute to labelling, but also to understand and interpret the errors and what they might mean in order to provide feedback. For example, we see problems with some of the label classification between author description and contribution (cf. Section 5.8.2). Our error analysis of the features enables us to see the annotator is most likely using contextual reference and missing surface clues are causing the classifier to make an error. Surface indications, however, are not always in the same sentence, and can be up to several sentences away. This understanding is important in giving feedback to the PG student as to why *LitCrit* may provide incorrect labels and what this may mean for their writing.

5.3 Features to Recognise Author Intention

An essential step in the classification task is feature selection and choosing what the best features should be. In this section, features are described and how they relate to features used by other authors of intention models. The approach taken here is motivated by previous work done in Argument Zoning Teufel (1999). This work provides the largest lexicon and pattern list of cue phrases within the Computational Linguistic domain, the main focus of this thesis. However, as highlighted in Section 2.2.3, Teufel herself says the work carried out in Argument Zoning labelling focuses on one sentence only and could better contextualise information between sentences. This was also a point highlighted by the authors of Research Writing Tutor, who say that sentence labelling could be improved by better consideration of information in preceding or subsequent sen-

tences (cf. Section 2.3.1.1). Our work is also focused on *Related Work* sections only whereas Teufel's list was developed on all sections of a research article. Therefore, the existing lexicon needs to be adapted to our specific section of a research article. We add to existing feature approaches by bringing in context between sentences using annotated co-reference chains to citations, the author's own paper and by using specific patterns that additionally include co-references and discourse relation markers. We show in ablation tests, in Section 5.7, these additions and modifications make significant improvements to the performance of our model.

5.3.1 Cue Phrases and Words

Almost all of the existing works on automating author intention and providing writer feedback have used cue words or phrases as part of their feature set to identify author intention. Studies of patterns in linguistic studies can be shown to date back as early as 1924 (Jespersen, 1924), and Biber (2006) argues that these patterns are not accidental, that phrases are consistently functional with their high frequency an indication of expected formulaic occurrence. This is seen in multiple studies in English for Academic Practices (EAP) that look at different aspects of these frequent phrases, such as those mentioned in Section 2.2.1.1 (Biber, 2006; Biber et al., 2004; Cortes, 2004).

Describing Swales' work in Section 2.2.1.1, phrases and words were shown to be linked to identifying author intentions in a sentence. For example, the intention of *establishing a territory* is shown to link to phrases, such as *it is well known that* or *previous research has shown*. Linguistic variation in cue phrases, however, can cause issues in recognising intentions. Teufel (1999) points to two types of cue phrases in her work in automating Argument Zoning. The first, she calls formulaic, which are relatively static syntactically, and the second type that has more syntactic variation. This variation occurs due to the many linguistic forms of expressing what she describes as *who-does-what*. For example, there are many different ways an author can express their work is a continuation of another. Table 5.1 shows different ways of expressing these variations of the author using the work of someone else. This variation in phrasing is harder to capture than more static expressions, and Teufel (1999) tries to overcome this using what she calls *Agents* and *Actions*. *Agents*, represent a form of attribution

Original Sentence	Replaced Sentence
We base our model on the work of	US_AGENT Action_USE US_AGENT on the work of
We use the framework	US_AGENT Action_USE the framework of
Our work is based on	US_AGENT is Action_USE on

Table 5.1: Example table of how Teufel's Action and Agent types work. Examples used are adapted from (Teufel, 1999) pg 102, Figure 3.14 - Variability of statements expressing research continuation.

of ownership, e.g. the entity taking the action *the author of the paper* or *our algorithm* would be categorised as *US_AGENT*, *the authors of a cited paper* or *their algorithm* would be categorised as *THEM_AGENT*. *Actions* are verbs classified into semantic classes. Words are replaced with their lexicon equivalent, shown in the Table 5.1, and pattern matching looks for combinations of *Actions* and *Agents* to assign label types.

Cue Phrases Approach

The largest list available of patterns containing cue phrases and words, with some constrained by PoS tags, was developed in a study of Computational Linguistics literature[1] (Teufel, 1999). These cue phrases/words have been curated to semantic categories or to align to rhetorical moves of Argument Zoning. Not all of the rhetorical moves in Argument Zoning apply to our model of intentions for *Related Work*. Therefore, like Jurgens et al. (2018), who works specifically on citation function, we start with the original list and adapt it. We do not implement the *Action* and *Agent* types described above. However, motivated by this idea, we implement an alternative version of this using co-references. This is described in the co-reference section (cf. Section 5.3.3).

We extract from the original list all words and phrases that align to semantic classes but not those that align to specific Argument Zone labels or Agent Action patterns. We separate these into five different lexicons which are based on Verbs, Adjectives, Nouns, Negation words and Phrases (single and multi-word) which are not PoS constrained. Pattern matching is added to detect plurals when they were not present. Describing each lexicon in more detail:

[1]This is made available at https://github.com/WING-NUS/RAZ

- **Adjectives** - the original list was amended to include 82 polar words/phrases from Athar (2011). His list was manually derived from citation sentences indicating sentiment when recognising citation function from papers in the ACL Anthology Network. Not all of these were adjectives and were added to the correct part-of-speech list. The adjective list identifies positive and negative adjectives, e.g. *advantageous* - positive adjective, *inaccurate* - negative adjective.
- **Verbs** - 16 semantic classes of verbs are present with associated words and phrases, e.g. *USE* - made use, utilises, *PROBLEM* - neglect, hinder, *CONTRAST* - differ, contrast conflict
- **Nouns** - 17 classes of nouns, e.g. *WORK_NOUN* - strategy, system, technique, *PROBLEM_NOUN* - absence, lack, shortcoming.
- **Negation words** - when negation occurs, it reverses the polarity of meaning, capturing this enables an understanding of what kind of evaluation the author is offering, positive or negative, e.g. not, neither, never.
- **Phrases** - commonly occurring phrases of single or multiple words and these were assigned to classes, e.g. *be different from* ->CONTRAST. Additional words and phrases were added to the original list studying the frequency of n-grams occurring within *Related Work* sections.

Each sentence was parsed, matching for words and phrases, constraining this to part-of-speech where appropriate and transformed according to lexicon entries. Examples of transformed sentences are shown in Table 5.2.

5.3.2 Discourse Relations

Often text becomes more coherent when its units, such as clauses and sentences, are analysed together to derive the high level structure and information. For example, Figures 5.1 and 5.2 show two different types of discourse relation markers. *However* signals to the reader that a relationship exists between the two sentences and the reader can expect a contrast or comparison. In the second example, *Firstly* and *For example* indicate to the reader a connection exists between the sentences and the topic continues. Section 2.2.1 pointed out that most of the previous work in recognition of author intentions within academic writing

Original Sentence	Replaced Sentence
However, the mismatching between complex structures across languages and the poor parsing accuracy of the parser will hinder structure alignment.	CONTRAST, the mismatching between NEG_ADJ structures across languages and the NEG_ADJ parsing COMPARISON_NOUN of the parser will PROBLEM structure alignment.
We further develop this idea with some new features, which leads to a new framework.	We ADDITIONAL SOLVE this idea with some NEW_ADJ features, which leads to a new WORK_NOUN.
Our work differs from previous approaches in two key respects.	Our WORK_NOUN CONTRAST from BEFORE_ADJ WORK_NOUN in two MAGNIFIER_ADJ respects.

Table 5.2: Examples of sentences from *Related Works* parsed using the lexicons for cue phrases and words. The table shows the original sentence on the left and the transformed sentence on the right after parsing.

is sentence based and does not consider relations between sentences to support better labelling. Several works (Cotos and Pendar, 2016; Teufel and Kan, 2009; Kirschner et al., 2015) suggested that understanding this context may provide better intention recognition. A limited number of previous works do consider referring expressions, such as pronouns, to link to previously mentioned citation work, this is discussed in the next section. Understanding the relationships between text segments though is not trivial as these are not always found in adjacent positions and they can be implicitly embedded (Green, 2017; Stab and Gurevych, 2014). Implicit relations are those inferred by the reader in the absence of a discourse connective.

Inclusion of discourse relations is shown to improve Argument Zoning labelling in (Lin et al., 2014). They build a discourse parser that automatically recognises discourse relations, based on Penn Discourse TreeBank (PDTB).[2] Lin et al. (2014) show that Argument Zone labels have relationships with specific discourse relation types, and automatically recognising this increases the classifier performance, e.g. the Argument Zone **CTR** – a contrast zone – is more likely to contain CONTRAST discourse connectives, e.g. however, but, in contrast. However, they raise the problem of ambiguity in connectives, also highlighted

[2]PDTB is a large-scale resource of annotated discourse relations, both explicit and implicit, and their arguments over the 1 million word Wall Street Journal (WSJ) Corpus

Paraphrase Acquisition work such as that by (Lin and Pantel, 2001; Pantel and Pennacchiotti, 2006; Szpektor et al., 2004) is not constrained to named entities, and by using dependency trees, avoids the locality problems of lexical methods. However, these approaches have so far achieved limited accuracy, and are therefore hard to use to augment existing NLP systems

Figure 5.1: Example of a discourse relation connective *However* in Related Work sentences.

Our approach differs from previous work in two important respects. Firstly, our ultimate goal is to develop an image annotation model that can cope with real-world images and noisy data sets. Our solution is to leverage the vast resource of images available on the web but also the fact that many of these images are implicitly annotated. For example, news articles often contain images whose captions can be thought of as annotations.

Figure 5.2: Example of a discourse relation connectives, *firstly and for example* in *Related Work* sentences.

by Litman (1996). Litman shows such discourse indicators can also be used for semantic purposes, and the problem is determining which use is in play. In the first example below, *further* is used in a semantic role, in the second example it is used in an elaboration role, i.e. it signals an expansion on a point made in the previous sentence.

- This result is *further* away from our desired output
- *Further* to this we considered...

Approach to Finding Discourse Relations

The approach for finding relations is based on explicit discourse markers only. While implicit relations exist, they are hard to identify automatically. First, a list of explicit discourse connectives is taken from the Penn Discourse Treebank (PDTB), 2.0 Annotation Manual, Appendix A (Prasad et al., 2008). Frequencies of these connectives are studied based on patterns within *Related Works*. Some of these patterns already existed within the cue phrase/word lexicon described previously. How these patterns occur at the start of a sentence before the first verb was considered and if they convey a relationship as a standalone word or in conjunction with other words as a phrase. Mid-sentence occurrence was also considered for some connectives, such as *but*, *while*. Table 5.3 presents the categories and examples of some words and phrases. Some of the phrases already existed or partially existed within the lexicon, and the discourse relation markers superseded these with the lexicon being updated with these entries.

5.3.3 Co-reference Resolution

Co-reference (anaphora) resolution is the use of an expression that depends specifically upon an antecedent expression. In this work, the interest is in referring expressions to a cited work or the author's own work. Often, these are also found through pronoun use, such as the work by Kim and Webber (2006) who distinguish between the pronoun *they* anaphorically, whether it refers to the authors of a cited paper, or whether it refers to an entity that is discussed in the paper. Other methods of identifying these links have been through deictic expressions, such as *their methods*, *this experiment*. Co-reference to cited work has also been explored through identifying noun based co-references, such as referring to a cited work by their algorithm name (Rösiger and Teufel, 2014). Figure 5.3 provides examples of co-references. In the first example, the referring expression *the authors* refers back to the citation, and in the second example, an associative noun phrase is used to refer to the cited work.

Discussed in Section 5.3.1, Teufel (1999) does attempt to capture references to the author's own work and to the work of others, calling these Agents (e.g. US_AGENT - our paper), using pattern matching. However, Teufel (1999) shows that these co-reference phrases can be ambiguous when taking this pattern match-

Category	Example Words or Phrases
EXAMPLE	*For example, for instance, specifically, to illustrate, firstly/secondly*
CAUSE_EFFECT	*Thus, because, hence, therefore*
CONTRAST_COMPARE	*Whilst, while, despite, nonetheless, however* - all occurring at the beginning of a sentence or before the first verb. There were some contrast connectives that we only marked when they occurred with a reference to the author's work *rather, opposed, instead* e.g *instead we do X, instead we use X*
ADDITIONAL	*Also, additionally, further, in addition*
TIME	*Before, earlier, recent after*
SIMILARITY	*Like/Unlike likewise, in the same way* - at the beginning of a sentence
CONTRAST_BUT	*But* - when it occurred at the beginning of a sentence indicated a contrast.
CONTRAST_BUTALAS	When *but* occurrs mid-sentence and is followed by negation, a problem word from the lexicon, or a negative adjective from the lexicon.
CONTRAST_BUTWEDIFFER	This was *but* mid sentence, followed by a reference to the author's work.
CONTRAST_WHILE	If *while* was found in the middle of a sentence in conjunction with a co-reference to the authors work or a citation

Table 5.3: Discourse relation categories on the left with explanation or examples of words and phrases on the right that are used to identify the categories.

SVM regression has recently been used by(Li et al. , 2007) for sentence ranking for general MDS . The authors calculated a similarity score for each sentence to the human summaries and then regress numeric features (e.g. , the centroid) from each sentence to this score.

MENE (Maximum Entropy Named Entity) (Borthwick,1999) was combined with Proteus(a hand-coded system), and came in fourth among all MUC-7 participants. MENE without Proteus, however, did not do very well and only achieved an F-measure of 84.22% (Borthwick,1999).

Figure 5.3: Examples of two types of co-reference linking, the first linking by a deictic phrase *the authors* the second by an associative noun phrase *MENE*.

ing approach. For example, does *this paper* mean the previously cited paper or is it referencing the author's work? In addition, there is no resolution of the pattern matched to the first mention of the entity, e.g. there is no alignment between *the authors* in Figure 5.3 to the first citing sentence. The approach taken here differs using the data set, described in Section 4.6 (Schäfer et al., 2012), that has co-referencing annotation included. These annotations are not just co-references but are directly linked to the initial citation entity when it first occurs in the *Related Work*. Specifically, this annotation captures:

- a previously mentioned cited paper, e.g. this could have originally been cited as Smith et al. 1999, and all subsequent mentions such as *their work, their paper, the model, their result* are marked as a co-reference to the original citation. It also marks associative noun phrases and links these to original citations.
- reference to the author's own work, i.e. their work in the paper not previous work is marked.

In addition to the existing co-reference annotations in the data set, it is necessary to identify mentions to multiple works. These are manually added for the work in this thesis. For example, a co-reference to multiple cited works, e.g. *these previously mentioned works above* is marked. Firstly, a rule based pattern match

Original Sentence	Parsed Sentence
”However, this method is not sufficient...”	CONTRAST COREF NOT POS_ADJ
Our problem is quite different from the above work.	OURCOREF PROBLEM_NOUN is quite CONTRAST_DIFF from the MCOREF.

Table 5.4: Examples of the original sentence on the left side and the resulting sentence on the right side after being parsed for lexicons of cue phrases and discourse relations, and then the co-reference annotations.

is taken to highlight where these phrases occur, and then these are then manually assessed and added.

We have three representation types for a co-reference (i) COREF for any co-reference to a citation (ii) OURCOREF for a co-reference to the author's own work (iii) MCOREF for a co-reference to multiple citations, e.g. *the methods above*.

These co-reference annotations provide two additions to the feature set. Firstly, the lexicons are modified to include cue words or phrases containing co-reference types. The Agent and Action patterns developed by (Teufel, 1999) are used as a starting point to do this. Many of the Agent/Action patterns are not relevant though as they align to AZ labels or patterns that occur outside *Related Work* sections. Additionally, combining the co-reference markers with discourse relations and other cue phrases, not just Action types is more effective. The second addition is that the co-references are linked from the initial reference and every subsequent mention. This allows the system to track what citation a co-reference in a sentence is referring. This is used in Section 6.3 when determining context between sentences to provide feedback.

The final lexicon consists of the three previously described features, cue phrases and words, discourse relation markers and co-references. Each sentence is parsed matching for words and phrases, constraining this to part-of-speech where appropriate. Table 5.4 shows examples of parsed sentences.

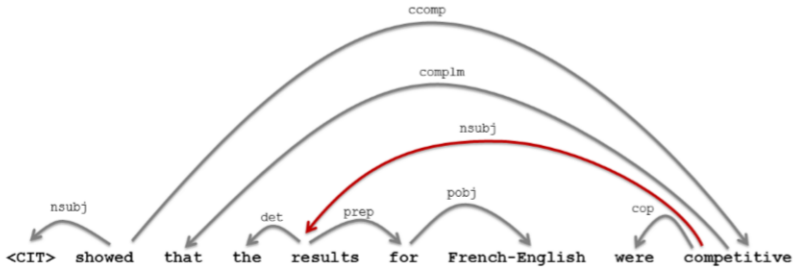


Figure 5.4: Dependency Structure Example

5.3.4 Citations Forms

Citation forms of integral and non-integral have been shown to be a contributing feature to author intention recognition (Swales, 1990), with studies of novice writers showing that they use a limited range of citation types (Thompson and Tribble, 2001).

A parser is implemented to identify three types of citation:

1. CIT1 – Those that form part of the syntax of the sentence (authorial)
2. CIT2 – Those that refer to the name of a system or known algorithm
3. CIT3 – Those that provide supporting evidence, found in parenthetical with no syntax

For example, the citation *"in (Smith, 1990)"* although in parenthesis would be of type CIT1 as it is part of the syntax, *"In recent years there has been a lot of interest in neural networks for question and answering (Smith, 2019; Jones, 2019)"* would be of type CIT3, and *"WNA (Johns, 2018) is probably the most widely used method ..."* would be of type CIT2.

This approach should help to discriminate between background sentences with citation evidence and citation description sentences.

5.3.5 Dependency Structures

Relations hold between lexical elements in a sentence. The grammatical relationship between words can be described by dependency structures, usually represented as triples **relationship(governor, dependant)** where the governor is the headword, and dependant is a dependant word (see the example sentence

Verb POS tag	POS tag description
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Table 5.5: Verb part of speech (POS) tab abbreviation on the left and the expanded description on the right.

in Figure 5.4). Each triple can also be thought of as a labelled edge from the governor to the dependent where the relation name is the edge label. Using dependency structures allows for capturing of long distance relationships between works (Athar, 2014). Consider the example sentence in Figure 5.4. The relation *nsubj* between the verb showed and CIT is captured as is the word *results* the nominal subject of the subordinate clause. Sentences with co-references are used to determine dependency structures and this is done for *nsubj* and *dobj* only.

5.3.6 Additional Verb Features

In addition to the verb lexicon described above, part of speech (POS) tags are used to identify verbs, treating the six possible VB tags (VB, VBD, VBG, VPN, VBP, VBZ) as binary features of being present or not in a sentence. Each POS tag and its description is described in Table 5.5.

5.3.7 N-grams

N-grams have been shown to perform well in NLP tasks. Liakata et al. (2012) show a 40% contribution to classifier results and Cotos and Pendar (2016) work is mainly based on n-gram features of 650 *Introductions*. The corpus in this work is much smaller (*Related Work* from 94 articles), nonetheless, n-grams are included as they will capture lexical phenomena that may have been overlooked by other features. Uni-grams, bi-grams and tri-grams occurring with a frequency of $\geq$5 were experimented with, and the final model uses bi-gram and tri-grams

only with no stop word removal.

5.3.8 Positional information

Positional information records where in a text a sentence occurs. Teufel (1999) used positional information including the absolute location of a sentence, section location and paragraph structure. Her work showed that relative sentence position was useful for identifying background sentences, as these are more likely to occur in an *Introduction* or *Related Work* than a *Results* section. Other works have also shown position information to have importance in sentence type identification, such as in citation function work (Jurgens et al., 2018), and in recognising CoreSC sentences (Liakata et al., 2012).

Initial experiments found that relative sentence location added no value. Other works look at whole articles, and relative sentence position would be useful for finding intentions linked to sections, but as a *Related Work* is only one section, this is the likely reason that this added no value. Instead a binary indicator for paragraph start and end sentence is used, manually added from the original PDF. This is similar to the feature in (Teufel and Moens, 2002) of paragraph structure. The expectation is this would work as a sentence relative position for this one section, as many background statements will come at the start of paragraphs, and towards the end of paragraphs, authors will be more likely to relate their own work.

5.3.9 Subject of Sentence

A sentence subject label is assigned to a sentence to decide if it is about a citation, background or field information, author's work, or a combination of author's work and cited work. This formed part of the annotation guidelines instructing annotators to decide this before applying a label (cf. Appendix C) and annotators highlighted this as an effective way to help decide on a label. This subject feature is based on rules, including sentence and previous sentence features of co-reference markers and paragraph start and end markers. There are six subjects: Background, Cited Work, Author's Work, Cited Work, Author's Work and Text.

5.3.10 Sentiment

Teufel et al. (2006b) work on automated recognition of citation function shows a strong relationship between function and sentiment. Each sentence is parsed for a count of positive and negative words using the polar list described previously and any additional positive/negative adjectives in the lexicon.

5.3.11 Counts

Counts of sentence words, nouns, adverbs, discourse connectives, citation and citation type, were included.

5.4 Classifier Methods Used

All models are trained using LibSVM (Chang and Lin, 2011) with a linear kernel and default settings. SVM's are known to be robust to over-fitting and perform well in document classification tasks when features are sparse and the set of them is large and does not assume statistical independence, making it a more suitable method when features may be overlapping or interdependent. We use a linear kernel as this can be easier to interpret and allows us to gain a better understanding of the features and their role in labelling sentences. Initially, experimentation was also carried out with decision trees methods. However, when tested for reliability in multiple iterations both Random Forest (Breiman, 2001) and C4.5 (Sumner et al., 2005) were not only consistently lower in performance (12%), but rare categories showed large variation (15%) between iterations, and in some instances, labels would not classify. This was likely due to feature overlap and some labels being multi-class. Due to the unreliability of its performance, the decision tree approach was not pursued further.

5.5 Label Distribution and Merging Infrequent Labels

When discussing previous work in Section 2.2.1 we highlighted the problem with sparse categories resulting in much lower prediction accuracy. Some cate-

Sentence Label	Count
BG(-+)	90
BG-NE	257
BG-EP	171
CW(-+)	133
CW-DESC	707
A-USE	59
A-DESC	107
TXT	21
A-CW	151
A-GAP	59
Total	1755

Table 5.6: Label class distribution of labels used in classifier.

Sentence Label	Count
BG(+)	17
BG(-)	73
CW(+)	21
CW(-)	112
CW-COM	24
A-USE	14
A-SIM	45

Table 5.7: Label class distribution of labels that were merged

gories are rare in the data used here particularly, BG(+) and CW(+). These two categories were collapsed with their corresponding BG(-) and CW(-) to create BG(+-) and CW(+-). Additionally, A-USE (author's work builds on/adapts/uses X), and A-SIM (author's work is similar to X) were merged into one category – A-USE. Finally, CW-COM (comparison of two cited works) was merged into CW-DESC (cited work description). The category of OTHER was particularly infrequent, and it was decided to re-annotate these. On reflection, these could be seen as background or author gap/description labels and were re-annotated to these categories. Table 5.6 shows the final distribution of the labels in the 94 *Related Work* sections and Table 5.7 the numbers for the merged labels. One *Related Work* was dropped from the previous chapter findings as it contained several OCR errors.

5.6 Experimental Setup and Evaluation

5.6.1 Baseline

Two baselines are provided, one with n-gram features only and one with all features based on the majority class.

5.6.2 Evaluation

This work is similar to other automated classifications but not directly comparable as schemas and experimental settings differ. The results are more comparable to the works of (Teufel, 1999; Jurgens et al., 2018; Teufel and Kan, 2009) as the same pattern list from Teufel (1999) is used as a starting point. These works use Naive Bayes, Random Forest and Maximum Entropy as classifier methods. Presented are the published Macro F1 scores, range of F1 scores for labels and the number of labels in the schema for comparison (Table 5.9). Also included are the results of Research Writing Tutor (Cotos and Pendar, 2016) (cf. Section 2.3.1.1) which focuses on writing feedback for *Introductions*. This is a much larger corpus using 650 annotated *Introductions* but fewer features, focusing on unigrams and trigrams. However, it also uses SVM for classification. Where available also reported is precision, recall and accuracy from these works to compare against this work's best performing model in Table 5.8.

Reliability of the model is important to ensure consistent results. Therefore, in addition to 10-fold cross validation, 10 iterations of the *All Features* model is carried out, reporting on mean precision, recall, F1, accuracy and variance in Table 5.8. Each iteration starts from a different seed. None of the iterations produced significantly different results, demonstrating reliability and low variation. Significance, where noted, is tested with corrected t-test, $p < 0.01$, (Nadeau and Bengio, 1999). Both precision and recall is important and we therefore focus on F1 scores when reporting and include Micro and Macro average for F1. Macro-averaging treats all classes equally and can be preferred if a model is to perform across all classes. Micro-averaging may be preferred if the density of a class reflects it importance (Jackson and Moulinier, 2002).

Features performance and influence on the label F1 scores is also reported with leave one out (LOO), which highlights the performance decrease when a single

feature is omitted and single features (SF), which highlights the contribution of a single feature to performance. Looking at individual label features is important as having just one label perform poorly, such as being able to recognise an author gap sentence or where an author says how their work is different, will impact the ability to give reliable feedback.

5.7 Results

5.7.1 Classifier Performance

Comparison of results to those mentioned in Section 5.5 is presented in Table 5.9. Comparing F1 scores overall, this work has better results than other systems by a reasonable margin. The range of F1 scores for the labels is also similar to other systems. This work produces better results than Research Writing Tutor (Cotos and Pendar, 2016) in F1 scores but not in overall accuracy. Their work is based on a bigger annotated corpus. The final *All features* model significantly outperforms both the baselines of n-grams and majority class. Re-running the classification (no novel features) removes the manual additions described in cue phrases and words, discourse relations, co-references and subject labels. This reverts back to the original pattern list by Teufel (1999) without any labels that specifically align to AZ labels. This results in lower performance significant p $<$0.01 than the *All features* and the majority baseline.

Features	**Precision**%	**Recall**%	**F1**%	**Accuracy**%
ALL	69 (0.50)	70 (0.40)	70 (0.50)	70.00 (0.48)
(Cotos and Pendar, 2016)	69	55	61	72.90
(Teufel and Kan, 2009)	48	38	41	66.80
height				

Table 5.8: Classifier performance and mean scores after 10 iterations with variance in brackets(%) for the work done in this thesis (All) and for the work of (Cotos and Pendar, 2016) and (Teufel and Kan, 2009).

System	F1/Range %	Number of Labels
(Teufel and Kan, 2009)	41 (19-81)	8
(Jurgens et al., 2018)	53	6
(Teufel, 1999)	68 (28-86)	12
(Cotos and Pendar, 2016)	61 (36-85)	17
Our Work		
-All features	70* (25 -88)	10
- no novel feat	54 (15-87)	
Baseline		
Ngram(B,T)	39 (2-68)	
Majority	57 (-)	

Table 5.9: Published results for the works used for comparison in the top of the table and the classifier results for work in this thesis in the bottom half of the table. The *All features* models is significantly better than no novel features and the two baselines, * significant 0.01

5.7.2 Feature Contribution

Feature contributions by single feature and leave one out are presented in Table 5.10. The top part of the table is leave one out and the lower part is single feature(s) for each category, the lowest score is in bold and in brackets are any scores higher than the *All features* model.

More frequently occurring categories, CW-DESC (cited work description) , BG-NE, BG-EP (background sentences with and without evidence) are more robust to feature omissions. Features are not independent, so many of the patterns cover the n-gram features, which may be why leaving out n-grams has less impact than expected. In the lower part of the table, n-grams as a single feature contributes most to labels TXT and CW-DESC. Compared to other works that used n-grams, the corpus used here is much smaller at $<$3000, whereas Liakata et al. (2012) used $\sim$42000 and Cotos and Pendar (2016) had $\sim$27000. It would be expected in a much larger corpus that n-grams will contribute more as a feature.

Sentiment contributes in a small way to performance but particularly in the evaluation labels, BG(+-) and CW(+-), as expected. Surprisingly, sentiment contributes to the text label. However, within text-labelled sentences, both of these counts are zero, which may explain why it contributes here.

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
ALL	39	72	73	53	84	48	47	88	63	25
Feat-(LOO)										
-subject	33	62	71	51	81	**49**	41	85	**64**	22
-n-grams	33	70	70	53	84	**50**	39	83	62	25
-verbtense	35	71	72	51	84	48	46	88	**66**	**32**
-sentiment	34	71	71	50	84	46	43	67	61	**28**
-counts	**40**	72	73	52	84	**50**	46	87	**64**	**26**
-Tot cit	38	71	**74**	**54**	**85**	**49**	**48**	88	**64**	**26**
-paragraph	**40**	71	73	**54**	84	**49**	47	87	62	22

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
ALL	39	72	73	53	84	48	47	88	63	25
Feat-(SF)										
-Allpatterns	30	54	**74**	41	77	**57**	**48**	80	**65**	**26**
-subject	-	58	-	-	80	-	45	75	46	-
-sub+patt+dep	31	72	73	47	83	**55**	46	84	63	**27**
-n-grams	11	31	21	24	62	23	04	68	39	02

Table 5.10: F-Measures (%) for features and labels, 10-fold cross validation, higher scores are in bold. The top half of the table is leave one out and the bottom half uses those features only.

Neither of the evaluations labels, BG(+-) or CW(+-), perform as well as expected. These two labels are merged from the annotation schema, positive and shortcoming/problem into one evaluation label. The original labels are both different linguistically, and this possibly proves more difficult for the classifier. We re-run the classifier, splitting the label into BG and CW (+) and (-) labels to investigate this. These labels suffer from sparseness (see Table 5.6 and 5.7). Re-running the classifier with the evaluation split results in a significant ($p < 0.01$) drop in the overall accuracy of the classifier to 66.30%. The F1 scores for the (+) labels are unacceptably low for prediction: BG(+) 9%, CW(+) 17%. Splitting these labels marginally lowers the F1 scores for CW(-) from 53% to 51% and BG(-) from 39% to 34%.

The removal of the paragraph start and end markers makes relatively little difference, except for the A-GAP (author gap) category. Being a rare category, this addition, although small is important. Total citation counts and counts of adverbs, words, nouns and discourse connectives seem to make the performance of the classifier worse on many of the labels, although not significantly so. There is an overlap in total citation counts with the count of citation types, perhaps indicating this feature could be omitted. Most categories are negatively impacted by the removal of the subject label with the exception of author A-USE (uses/build/similar to cited work) and A-CW (authors work differs from cited work). The features added to the pattern list, dependencies and subject label are very close to the performance of the *All Features* model. Performance improves on the rare label A-GAP (author gap) with just these features alone.

As a single feature, subject is important to the classifier performance and contributes to several of the labels: BG-NE (background with no evidence), CW-DESC (cited work description), A-DESC (author description) and A-CW (author and cited work differ). Leaving out subject label was the only feature to cause a drop in classifier performance that was significant. In Table 5.11 and Table 5.12, experiments from using a gold subject label and using a history feature of the previous label are presented. History label was previously shown by Liakata et al. (2012) to contribute to sentence classification. The gold subject label was determined from the annotated label. Determining this label accurately has an almost 15% increase in the performance of the classifier and an increase in F1 score for all label categories. This increase is significant ($p < 0.01$). It should be

noted that A-GAP individual F1 score is still low at 40%. Including a previous label also increases the classifier performance, but this increase was not a significant increase. It does, however, increase some of the individual F1 scores as highlighted in brackets, particularly the evaluative labels. However, the F1 score for TXT drops from 88% to 63%.

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
ALL	39	72	73	53	84	48	47	88	63	25
+Prevlabel	**50**	60	70	**60**	**86**	**51**	46	63	61	**27**
+GoldSubj	**61**	**86**	**88**	**67**	**94**	**68**	**72**	**100**	**88**	**40**

Table 5.11: F-Measures (%) for labels using all features and all features with gold subject and previous label. Bold indicates results are higher than the original *All features* model.

5.8 Mis-classification Error Analysis

There are several reasons mis-classification could occur, e.g. errors in the original parsing of the data, phrases/words that have not been encountered before. The classifier reports error numbers and the confusion matrix can be used to determine where the labels are mis-classified to. However, this does not give any indication as to why these labels where mis-classified. This section discusses the manual review undertaken of errors in intention labelling and how these relate to features.

The error classification rate for the best classifier in the previous section is just

Features	Prec%	Recall%	MicroF1%	Macro F1%	Acc%
ALL	69	70	70	60	70.00
+Previouslabel	71	72	71	60	71.72
+GoldSubject	84	85	84	76	84.60*

Table 5.12: Classifier performance with mean scores after 10 iterations comparing the all features and then *All features* model adding the previous label feature and *All features* adding the gold subject label. GoldSubject is significantly better than the previous All features model, * significant 0.01

under 30%. Five runs of the experiment using different seeds of classification are undertaken, and wrongly classified labels are compared. This is not an exact comparison due to the nature of the experiment resulting in different sentences in test sets, but approximately 90% of the mis-classified labels occur in all five runs. This accounts for 25% of sentences within the data-set. These labels were chosen for the error analysis study. Each sentence along with its features and the label it was mis-classified to most frequently was extracted. Error analysis was performed manually to assess if there was any consistency between feature types and mis-classified labels. Three re-occurring error types were found:

1. Data errors (4%), e.g. missed co-references, wrong annotation labels
2. Ambiguous or multi-label issue (5%), e.g. sentences that are unclear as to whether they are background or cited works.
3. Linguistic – e.g. missing terms in lexicons (15%)

5.8.1 Data Errors

Several error types were found in the data:

- Citations had been parsed as citation type 1 (CIT1) instead of type 3 (CIT3) or vice versa.
- OCR errors caused problems in the sentence
- Co-reference mistakes or missing co-reference from the original annotation
- Annotator mistakes, e.g. missed a citation was present or missed the reference to the author's work.

One of the aspects noticed is differences in how co-references are annotated. Some annotators pick whole phrases and several times associative noun co-references were missing. While the annotation carried out is not done as part of this work, in the next chapter, when Computer Graphics papers are annotated it is important that consistency is applied.

After manually fixing all mistakes, the classifier is re-run resulting in a 1% overall performance improvement which is not significant with relatively little change in any individual intention labels.

> Hand-coded descriptions of body posture shifts and eye gaze behaviour have been show to correlate with topic and turn boundaries in task orientated dialogue (Cassell et al., 2001)

Figure 5.5: Example of a sentence that could be labelled as either CW-DESC or BG-EP.

> Firstly, our ultimate goal is to develop an image annotation model that can cope with real-world images and noisy data sets.)
>
> This approach allowed us to perform a systematic feature analysis on a large-scale real-world corpus and a comprehensive feature set.

Figure 5.6: Examples of sentences the annotator labelled as A-GAP but the system labelled as A-DESC.

5.8.2 Ambiguous Labels or Multi-labels

Ambiguous Labels

Observing the mis-classified labels, ambiguity existed in that the annotator label or the classifier label could both be potentially correct. This was often due to a lack of clarity in the writing through the use of a passive voice or through a misleading citation type. In the example in Figure 5.5, the annotator chose CW-DESC but the classifier labels this as BG-EP. It could be argued that the classifier is correct as it is unclear from the author's citation if it is one example or are they describing that specific work.

Another common error was between A-GAP and A-DESC. In the examples in Figure 5.6, the annotator marks this as A-GAP, but the classifier labels it as A-DESC. These sentences were probably by far the most subjective and highlighted earlier in Section 4.8.3 as being difficult for an annotator, which may be even more difficult for a classifier to get right. Although annotators are instructed to only mark labels when linguistic clues exist, this can be a problem, and domain knowledge sometimes means annotators will mark labels as A-GAP although

linguistic clues may not be present.

Multi-label When deciding on the sentence annotation unit in Section 4.2, it was highlighted that multi-labels could potentially be a problem, particularly when this came to classifying. However, reviewing the errors, there were only ten sentences that could have really be described as multi-label. These were long sentences that were missing punctuation or they were sentences separated into two sub-clauses by the use of **;**. Possible solutions to this would be to implement a parser with PoS tags to identify and label sentences separated by a **;** or to use a tool such as **Grammarly** to highlight problems with punctuation and warn that labels may be mis-applied in these instances.

5.8.3 Linguistic Clues Missing

Missing words in the lexicon seemed to mainly impact the correct recognition of evaluation sentences. Almost 130 sentences were missing a feature that identified a cue word or phrase related to evaluation. Positives were the most difficult for the classifier to identify but the number of positive examples are very sparse. These were sometimes difficult for annotators to pick out.

Recognising comparisons between citations and the author's work could also be challenging for the classifier as often these would start with an initial sentence that only stated the author's work was different, e.g. *Our work is different*. Then, the author would give two to three sentences explaining why their work was different. The annotator would mark these as A-CW (author-cited work differs), but by the second or third sentence, the classifier would mark this as A-DESC as no linguistic clues existed to pick up that this was a comparison and thus it just looks like author description.

5.9 Improving the Classifier After Error Analysis

5.9.1 Improving Subject of a Sentence Feature

Currently, the feature of sentence subject, described in Section 5.3.9, is the only feature when left out that caused a significant drop in classifier performance but an almost 15% increase (significant) was shown to occur using a gold sentence

Label	Total	Orig Guess	New Guess
BG	514	26%	19%
CW	838	13%	6%
A	167	24%	19%
A/CW	215	48%	53%

Table 5.13: Subject error improvement with the new subject feature method. In the left column is the subject label, the second column the total number of subject sentences, the third column the original error percentage and finally the new error percentage for each subject label.

subject label. The sentence subject is based on rules that assign a subject to a sentence. In this section, we investigate improving the rule assignment by incorporating more information from a preceding or following sentence. This will help in sentences where there is no linguistic clue as to the subject, e.g. those that carry on describing a cited work but there is no co-reference to signal this. The current subject feature, when compared to the gold standard, has an accuracy of 77%. The rules to determine the subject use co-reference markers and paragraph start and end markers.

To improve the accuracy of the subject feature, discourse marker relations are added to the rules. These are used to help determine that a sentence continues to talk about the same cited work or author's work, e.g. *for example, however, therefore*. This addition improved the subject feature in overall accuracy by 6% from 77% to 83%. Individual label increases are shown in Table 5.13. The best performing classifier from Section 5.7.2 with the *previous label* feature included is re-run using the new subject feature. This results in a significant increase in performance of the classifier ($p < 0.01$) to 76.28%, seen in Table 5.14. Additionally, this results in better performance for some individual labels (Table 5.15). The two BG labels of BG-EP and BG-NE in particular increase as does the CW(+-). There is a small increase in A-GAP but unfortunately, BG(+-) and A-CW drop. Referring back to Table 5.13, the label A-CW dropping in performance is not surprising as it is the only label whose error rate increases (marginally) with this new subject feature.

Features	Precision	Recall	MicroF1	Macro F1	Accuracy
+OrigPlabel	71	72	71	60	71.72
+NewSubjectGuess	76	76	76	64	76.28*

Table 5.14: Classifier performance (%) for the original features including previous label and using the new subject feature, * indicating significant, p $<$0.01.

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
+OrigPlabel	50	60	70	60	86	51	46	63	61	27
NewSubjectGuess	45	**78**	**81**	**69**	**90**	**60**	**65**	**65**	56	**33**

Table 5.15: New F1-Measures (%) for labels using the new subject feature compared to the previous results using *All features* and previous label. Increases are denoted by bold.

5.9.2 Adding a Label Suggestion Feature

The approach to features taken so far is a bag of words for the cue phrase features or discourse relations. Words and phrases are identified and replaced with their lexicon category, e.g. *In contrast* would be replaced with CONTRAST, or *in contrast to our work* is replaced with CONTRAST_WEDIFFER. These cue phrases though are either present or not, and no consideration is given to the order they occur in a sentence. In the original work on Argument Zoning, Teufel (1999) included formulaic patterns that considered orderings of the cue phrase/words. She studied patterns of these in order to assign Argument Zone labels or further semantic classes. These formulaic patterns are not necessarily relevant to this work as they mostly align to different author intentions to the ones in this thesis and to constructs that occur outside a *Related Work*. A similar approach to Teufel was taken in the work of Jurgens et al. (2018) on citation function. Their implementation used bootstrapping to learn over four times the manually curated patterns of Teufel.

The approach here is manual, like Teufel, and based on studying the sentences within the *Related Works*. In addition, we considered the co-reference, discourse relations and the gold subject labels. Patterns are constructed to capture the labels, examples of which are in Table 5.16. It was found that predicting if the

Semantic Representation Within Sentence	Label Suggestion
CONTRAST_MIDWHILE + Subject=A	A-CW
CONTINUE or SIMILAR + Subject = A/CW	A-USE
CONTRAST(*ANY) + PROBLEM or NEG_ADJ	EVAL
CONTRAST_BUTALAS + Subject=BG	EVAL

Table 5.16: Examples of rules that suggest the label of a sentences based on cue phrase substitution in sentences, co-reference, discourse markers and gold subject labels.

sentence was evaluative (suggesting label of EVAL) over whether it was BG(+-) or CW(+-) – background or cited work evaluation – was more effective. These labels are not used as the predicted label directly but fed to the classifier as an additional feature we call *Label Suggestion*. The best performing classifier with the new subject feature, described in the previous section 5.9.1, is re-run with this added feature.

A limitation with this approach, though is that these patterns are likely to be very discipline specific and evidence of this is presented in Chapter 7. Chapter 7 applies these patterns when predicting intentions in *Related Works* from the Computer Graphics discipline, and this addition makes the classifier perform worse.

Results - Adding Formulaic Label Suggesting Patterns

This additional feature to suggest an author intention label results in a very marginal but not significant ($p < 0.01$) increase to the performance with overall accuracy now reaching 76.34% – compared to 76.28% previously. There are very marginal increases in F1 scores presented in Table 5.17 with some marginal decreases. The biggest positive changes are in the F1 scores for TXT and A-USE. Overall, the manual approach to finding these label suggestion patterns is not sufficient enough, and this would benefit from a more automated approach, e.g. Jurgens et al. (2018) who use bootstrapping to learn patterns.

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
NewSubjectGuess	45	78	81	69	90	60	65	65	56	33
+SuggestLabelFeature	44	78	81	68	90	**64**	65	**71**	54	**34**

Table 5.17: New F1-Measures (%) results for labels using the labels suggestion feature comparing this to the results in the previous section of new subject guess. Increases denoted in bold.

5.10 Discussion and Conclusions

The manually annotated data set curated in the previous chapter is used to classify the author intention labels for *Related Work* feedback showing that this approach can achieve similar, and in some cases better, results than classifiers for other intention models. Overall, the automated recognition is good at 76.34%, almost reaching the human annotation agreement of 77%. The features used are described and these are related to existing works, with novel features, such as using co-reference specific to *Related Works* and the adapted pattern sets explained. The introduction of these features over and above the original pattern features from Teufel (1999) was shown to be a contributing factor to the performance of the classifier. This highlights the importance of understanding the author intentions of interest and looking for patterns that are specific to these. This is also a limitation in that these patterns are built on one section and within the Computational Linguistic domain. In Chapter 7, we explore how the classifier performs in another domain, Computer Graphics, and we see evidence of how discipline specific features are impacts classifier performance.

Whilst prediction is good in some categories, not all categories are good with A-GAP being one of the worst performing. Evidence shows that annotators are using knowledge to infer A-GAP and that linguistic clues are not present, making the job of a classifier particularly difficult. For the overall intentions of giving feedback, it may be prudent to understand the value for this category in feedback before preceding with trying to improve the classifier performance. It may be satisfactory to label the majority of these A-DESC and only use A-GAP when very evident linguistic clues are present, e.g. *Our contribution is, the novelty of our work is that ...* However, as we see in Chapter 7, when we look at using the classifier in the domain of Computer Graphics, who have more

A-GAP labels, combining data from two domains does improve the F1 scores for this category. This suggests this low score stems from a data sparsity issue and could, to some extent, be overcome with more examples of this type.

Error analysis reveals the importance of annotated data being correct but also highlights the challenges in creating a fully automated system that can reliably identify co-references. Whilst the task of co-reference is not covered in this thesis, it will need to be carefully explored if it is to be integrated in the future into the work done here.

Having access to more data examples would no doubt improve the ability of the classifier and allow the expansion of cue phrases to identify evaluative statements. However, access to pre-labelled data is limited due to the expense of manual annotation. Other works which use a feature-engineered approach have expanded the lexicon vocabularies using bootstrapping methods to identify semantically similar variants of phrases, used word embeddings or using words found to be semantically similar, e.g. using WordNet. The idea of bootstrapping (Abdalla and Teufel, 2006) is that is will have more power to generalise capturing linguistic variation for semi-fixed phrases. Jurgens et al. (2018) uses bootstrapping to automatically identify patterns that occur in text manually labelled with a citation function. These patterns consisted of lexical categories, PoS wild cards or tokens directly. Heffernan and Teufel (2018), who look for different wordings of problem and solution descriptions in scientific text, use Word2Vec (Mikolov et al., 2013) to identify semantically similar words using PubMed articles as their search expansion. The final selection of their word expansion sets is chosen manually. Both these works, however, are focused on specific domains and given that disciplines are known to favour specific vocabulary, there may be limitations as to how these would transfer to different domains. Recent work of Asadi et al. (2019) show that using WordNet roots for Nouns, e.g where nouns are taken to their more general form (e.g., *mm* and *cm* become *quantity*), is a useful feature for author intention identification. This type of application of WordNet is one possible avenue that may assist in transitioning the pattern list to another domain by making words generic.

Chapter 6

Visualising the Related Work Narrative

6.1 Introduction

In this chapter we carry out a further study, using the same material as our first study, but this time presenting the material within **LitCrit** the writing analytic tool developed as part of this thesis work. We show how this approach changes the PG students' thinking and perceptions of the *Related Work*, bringing them more in line with experts. Also described is how discourse segmentation is carried out to provide overall feedback on the *Related Work*, not just the highlighting of author intentions. This chapter also compares how accurately the automated system segments and labels text for feedback compared to a human.

6.2 LitCrit Interface Design

Design of systems and how humans interact with such designs is a specialised field. With reference to building writing analytic writing tools, work has been undertaken to understand and evaluate how best to present feedback to students (Cotos, 2009; Shum et al., 2016; Gibson et al., 2017). These aspects of study are outside the scope of this thesis, but instead, we borrow from what others have learned in such design. In particular, we turn to the writing analytic tools mentioned previously, AcaWriter and Research Writing tutor (cf. Section

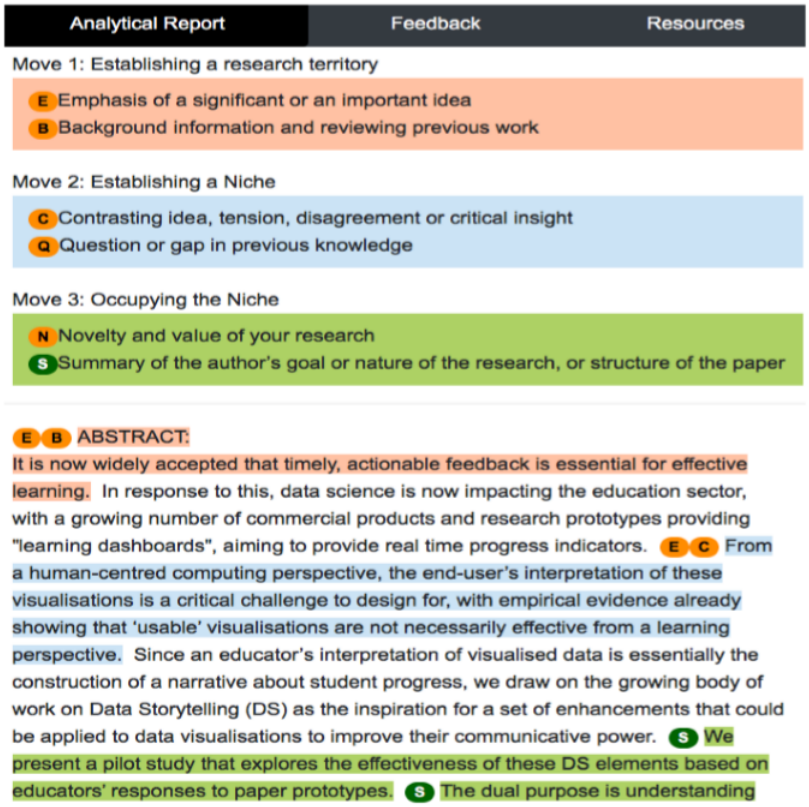


Figure 6.1: AcaWriter CARS Parser from (Abel et al., 2018)

2.3.1.1). These highlight narrative structure with intentions providing additional feedback comments or suggestions on what the highlighting may mean at the side. O'Rourke and Calvo (2009) argue that visualisation can provide insight into latent features in writing, mitigating some of the problems of subjectivity. Highlighting is achieved through the use of tags and colours, which pertain to the author intentions. For example, AcaWriter in Figure 6.1. Tags are usually colour coded with segments of the text coloured to indicate the tag they belong to. Gibson et al. (2017) argue that coloured symbols allow for the instant recognition of key elements. This design approach is used with pharmacy students in reflective feedback (Lucas et al., 2018). Their approach to design is successful, and early signs show that students find feedback presented actionable. Similarly, Teufel et al. (2009) finds evidence that the highlighting of Argument Zones allows students to interpret the major points of the article better. However, student responses also indicate that some preferred the version of text without highlighting and to apply highlights on demand.

The design of **LitCrit** takes the same approach of tags and colour coding seg-

ments of text. Author intentions in the narrative are highlighted using colours, linking to a key on the side that provides further explanation. A comment box provides feedback about intentions that are missing and present, thus bringing the PG student's attention to suggestions of how their writing could be improved.

Pilot feedback A pilot study was conducted using five PhD students not linked to any of our user studies who were studying within the Computational Linguistic discipline. These participants were asked through the author's network. The pilot study was informal and presented the seven *Related Works* within **LitCrit** to each participant through the web interface, and they could click through each in turn. Each participant was asked verbally for feedback about what they liked, did not like or any aspects that confused them. This initial version of **LitCrit** presented each author intention sentence colour-coded differently and a label key down the side, linking the colour to the author intention label. Feedback highlighted that it would be better to place the author intention labels at the beginning of each sentence because it was too confusing to move between the label key at the side and the different coloured sentences. Feedback also suggested to colour-code sentences to the main category they belonged (Background, Cited Work, Author) as opposed to individual colours for each author intention label. Too many colours was found to be confusing.

Figure 6.2 presents a screen-shot of the **LitCrit** interface after pilot feedback was implemented. When the feedback tab is activated, the *Related Work* is highlighted with the author intention labels, a key to labels and colours at the right side and the feedback comments in the box below the *Related Work* text. When the feedback tab is not activated the *Related Work* is presented in plain text.

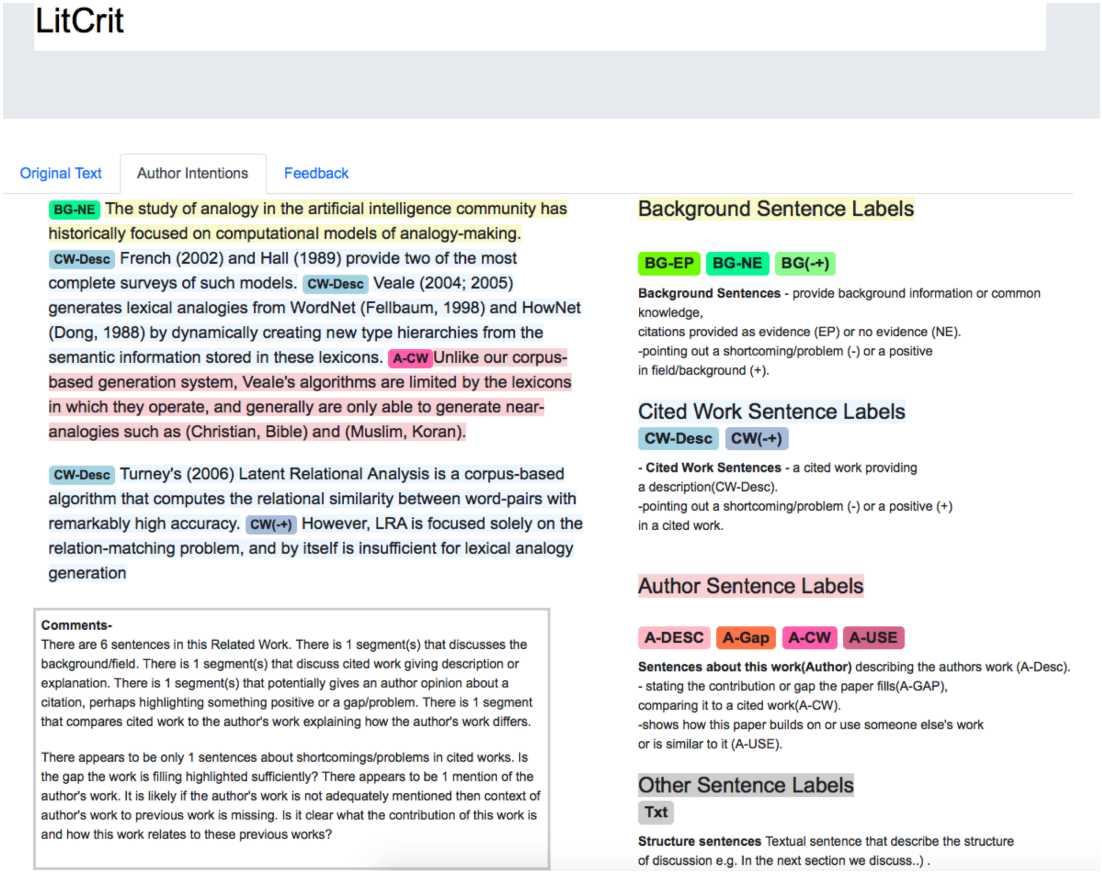


Figure 6.2: Screenshot of the interface of **LitCrit** with author intention labelling highlighted and feedback present.

6.3 Generating Author Intention Labels and Feedback for LitCrit

In the study described in Section 6.4, author intention labels and feedback are semi-automated. The focus is on assessing if the approach impacts the student responses and, therefore, any ambiguity is eliminated in the evaluation that may be introduced from an automated approach, e.g. if a wrong label or incorrect feedback is given. Whilst this has the advantage of not introducing ambiguity, it has the potential to make the evaluation of *LitCrit* more optimistic, particularly when participants are asked to evaluate *LitCrit*. The author intention labels are initially labelled using the best classifier, described in Section 5.9.2, and then manually corrected. This section describes the automated approach to segmenting the text and providing feedback (Comments box in Figure 6.2).

6.3.1 Discourse Segmentation for Author Feedback

The author intention labelling works at the sentence level, but discussion about a single work or author's work often extends over several sentences. Capturing these sentences into a contiguous block is essential for generating feedback. Using just the author intention labels, we cannot distinguish between a sequence of sentences each discussing a different *cited work* and a sequence of sentences all discussing the same work. To make this distinction we need to be able to, for example, identify multi-sentence segments in which the same *cited work* is discussed or identify multi-segment sentences that discuss only the author's work. Taking the sentences below as an example, this means we would want to create a segment after the first sentence, as this discusses one particular cited work. The next segment would include sentence 2 and 3 as these discuss the same cited work.

CW-DESC Ferrandez and Peral (2000) proposed a hand-engineered rule-based approach to identify and resolve zero pronouns that are in the subject grammatical position in Spanish

CW-DESC In Iida et al. (2006), they proposed a machine learning approach to resolve zero pronouns in Japanese using syntactic patterns.

CW(-+) Their system also did not perform zero pronoun identification, and assumed that correctly identified zero pronouns were given as input to their system.

Splitting text like this is known as discourse segmentation. Different approaches to discourse segmentation have been taken depending on the informational needs as to whether the text can be considered sequential or hierarchical, or if relationships need to be determined between text segments. Previous work has considered segmenting sentences within a document based on topics, such as TextTiling (Hearst, 1997), which tries to find local topic discussions within a text using repetitive terms and those that are closest in meaning. Other works have focused more on dividing the text into structural or functional roles, e.g. cause, elaboration based on theories such as RST, e.g. (Afantenos et al., 2010; Sporleder and Lapata, 2005). Capturing segments of discourse related to the same subject though is challenging as these are not always found in adjacent positions and may be implicitly embedded in text (Green, 2017; Stab and Gurevych, 2014).

6.3.2 Discourse Segmentation and Feedback Approach

Our coding of linguistic features to enable segmentation is inspired by Passonneau and Litman (1997) whose work, although on utterances, uses features, such as Noun Phrases and their co-references, and cue phrases, such as discourse relation markers. Our segmentation of the text is based on grouping sentences into contiguous segments that discuss a cited work, author's work, compare a cited work and the author's work or discuss the background/field. The segments defined match those of the subject feature, described in Section 5.3.9 and found in Table 6.1. Every segment starts at the beginning of one sentence and ends at the end of some subsequent sentence. No segment ends in the middle of a sentence.

6.3.2.1 Segmentation Process

Each *Related Work* is initially segmented using paragraph markers from the original paper. Following this discourse relation markers, described in Section 5.3.2, are used to mark sentences for linking. Discourse markers are useful when no co-reference exists. An example can be found in Figure 6.3. The last sentence starts with *Furthermore* and indicates a continuation of the claim or argument

Segment Types
Background
Citation
Author
Citation and Author
Text

Table 6.1: Segmentation types used to separate the *Related Work* discourse

This paper proposes a dynamic context- sensitive tree span trying to cover necessary structured information and a context-sensitive convolution tree kernel. Furthermore a composite kernel is applied to combine our tree kernel and a state of the art linear kernel..

Figure 6.3: Example of discourse connective link

started in the previous sentence. Where discourse relation markers occur before the first verb of the sentence, the sentence is marked to create a potential link to the previous sentence.

The sentences are then parsed in sequence and a segment break is added when (i) a new citation occurs with no co-reference to a preceding sentence or no discourse marker, and this is not a background sentence; (ii) a sentence is about the author's work only, i.e. no comparison to a cited work, and the previous sentence was: cited works, background, TXT; (iii) a sentence is a background sentence, and the previous sentence was: author, cited works, TXT; (iv) the sentence is a TXT sentence and the previous sentence was not TXT.

6.3.2.2 Segment Labelling and Generating Feedback

There is a need to generate overall feedback based on the segments within the *Related Work*. This feedback is generated based on labels applied to each segment. Segments labels are applied automatically based on the author intention labels that occur within the segment. Labels for segments can be seen in Table 6.2. For example, if a segment only contained citation explanation it would be

Segment Labels
Author's work - description only
Author's work - highlights novelty
Author's work - based on/uses or similar to other work
Author's work - author's work differs to other work
Background - with evidence
Background - without evidence
Background - with insight highlighting a positive or negative(gap or problem)
Cited Work - description only
Cited Work - with insight highlighting a positive or negative(gap or problem)
Comparison - author and cited work
Text - structure information to article No Label- failed to find a label

Table 6.2: Segmentation labels used to label segments for feedback

labelled *Citation - description only* or if the segment contained citation explanation and a CW(-+) label it would be labelled *Citation - with insight highlighting a positive or negative(gap or problem)*.

The generation of overall feedback takes a template approach reporting on the number of sentences there are in total, then on how many segments are present for each label in 6.2. Finally, if there is only one mention of the author's own work, it suggests that the context of previous work to the author's work may be missing and the contribution may not be clear. If there are two or fewer insight segments, it suggests that critical evaluation may be missing and the gap or problem the author is filling may not be clear. Examples of *Related Work* and their generated feedback can be found in Figures 6.4 and 6.5.

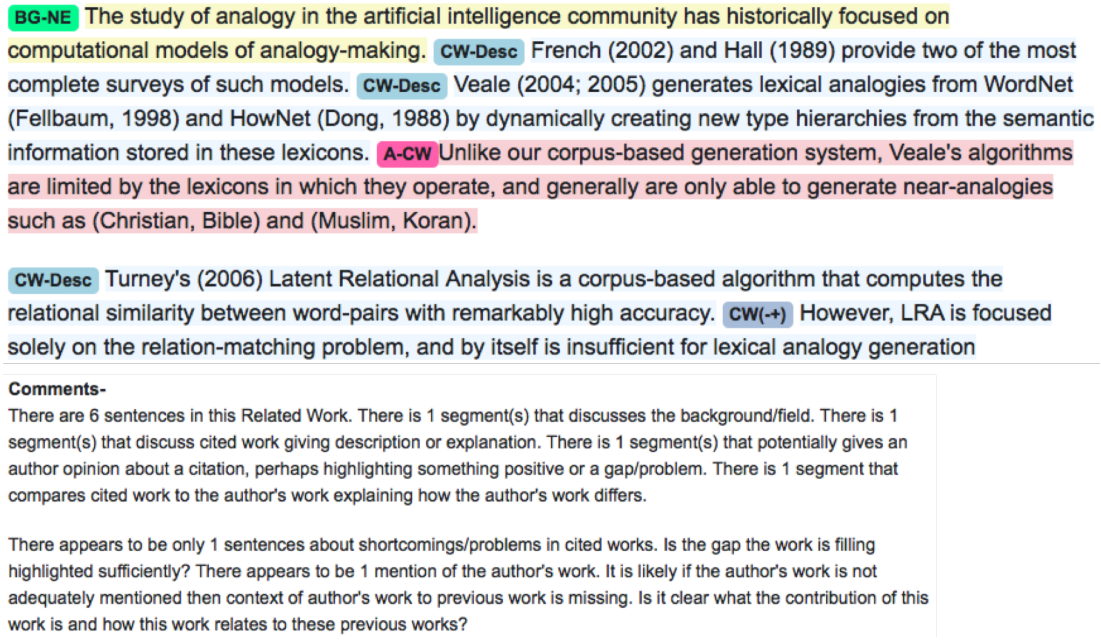


Figure 6.4: First example showing *Related Work* highlighted with author intention and feedback (Comments box) that is generated within **LitCrit**

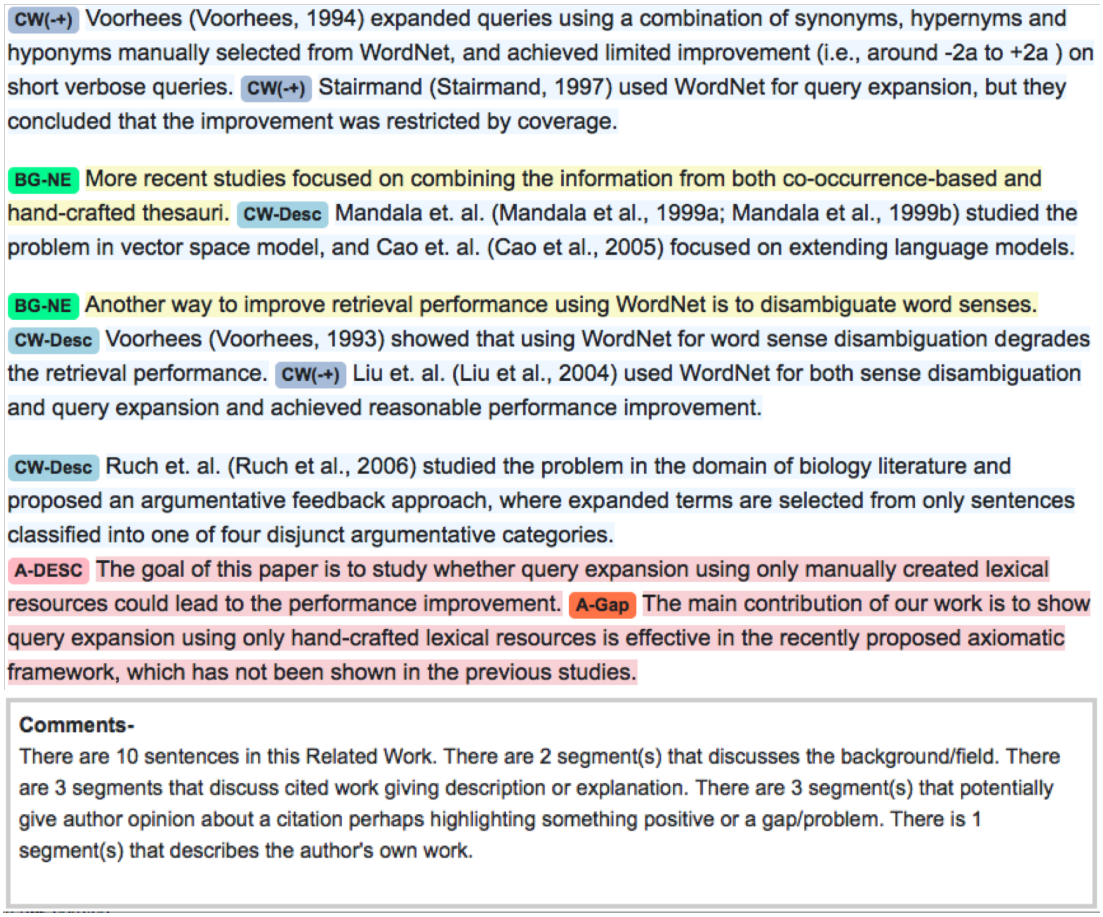


Figure 6.5: Second example showing *Related Work* highlighted with author intention and feedback (Comments box) that is generated within **LitCrit**

6.3.3 Accuracy of Segmentation and Segment Labels

The user study, described in Section 6.4, manually corrects the feedback generated, but the study does evaluate the usefulness of this feedback from the user's perspective. The goal of the evaluation in this section is to measure how accurate the automated system is at determining segments and segment labels.

Using segments and labels marked by a human subject, the evaluation compares (i) How many segments the automated system gets correct, and (ii) For the segments it gets correct, how many of these agree on segment labels. In addition, we discuss the types of problems that result in errors.

6.3.3.1 Experimental Procedure

A human subject is required to carry out manual segmentation of the *Related Work* section for comparison to the automatically generated segments. These segment boundaries are based on finding blocks of sentences that align to segments as described in Table 6.1. The segment boundaries are relatively hard and not subjective, thus, it was considered acceptable to use the author of this thesis as the human subject for evaluation.

Each *Related Work* is put into a file with newline breaks between paragraphs, as per the original *Related Work*. Each file is read by the subject and newline breaks added to indicate segment boundaries. These files are then processed into an Excel file with one row per segment. Following this, the subject labels each segment, selecting the appropriate label (Table 6.2) from a drop-down list in Excel.

6.3.3.2 Evaluation of Segments and Labelling

There are existing measures of segmentation with one of the most popular methods being WindowDiff (Pevzner and Hearst, 2002). WindowDiff compares the portion of segment boundaries from the ground truth compared to the algorithm boundaries using a sliding window. However, this method has been shown to emphasise mistakes at the beginning and end of the text. Errors near the beginning or end of a segment are counted slightly less than other errors (Lamprier et al.,

Segmentation	Score
Precision	91
Recall	90
False Positive Rate	9

Table 6.3: Precision, recall and false positive rate of the system compared to a human in discourse segmentation.

2007). Overall our approach to segmentation is simplistic, and our evaluation is based on a straight forward accuracy measure.

In evaluation, the human segmentation is taken as ground truth. The segments from the system are counted to compare (i) how many segments agree with the human segmentation and we report Precision, Recall and False Positive Rate, calculated from equations (6.1), (6.2), (6.3). (ii) for the segments that agree how many labels agree. We also review the errors that occurred.

$$Precision = \frac{SegmentsCorrect}{No.Segments\,found\,by\,system} \tag{6.1}$$

$$Recall = \frac{SegmentsCorrect}{Actual\,Number\,of\,Segments} \tag{6.2}$$

$$False\,Positive\,Rate = \frac{SegmentsWrong}{Actual\,Number\,of\,Segments} \tag{6.3}$$

6.3.3.3 Results - Segmentation and Segment Labelling

Overall the segmentation works well, reporting precision at 91%, recall at 90% and a false positive rate of 9%. (Table 6.3). The system overall produced more segmentation than humans, 836 compared to 830. Table 6.4 shows an example of a segmented *Related Work* with label segments.

The system was incorrect on 75 segments and we reviewed each of these to determine why segmentation went wrong. In 75% of cases the problem was a result of data error. Five of these were due to incorrect author intention labels, but the majority were either missing co-references or citation parsing that went wrong. Particular problems with citation parsing are: where names are used in the sentence, but proper citation form is not, e.g. *Hovy finds that ...*, rather than *Hovy and Dirk (1999) finds that ...*, or when the parser thinks it is a citation of

Most of the previous works conduct structure alignment with complex, hierarchical structures, such as phrase structures (e.g., Kaji, Kida & Morimoto,1992), or dependency structures (e.g., Matsumoto et al. 1993;Grishman ,1994;Meyers, Yanharber & Grishman 1996;Watanabe, Kurohashi & Aramaki 2000). However, the mismatching between complex structures across languages and the poor parsing accuracy of the parser will hinder structure alignment algorithms from working out high accuracy results.
Segment Label : *Background - with insight highlighting a negative or positive(gap or problem)*

A straightforward strategy for structure alignment is parse-to-parse matching, which regards the parsing and alignment as two separate and successive procedures . First, parsing is conducted on each language, respectively. Then the correspondent structures in different languages are aligned (e.g., Kaji, Kida & Morimoto 1992; Matsumoto et al. 1993;Grishman 1994;Meyers, Yanharber & Grishman 1996; Watanabe, Kurohashi & Aramaki 2000).
Segment Label : *Background - with evidence*

Unfortunately, automatic parse-to-parse matching has some weaknesses as described in Wu(2000). For example, grammar inconsistency exists across languages; and it is hard to handle multiple alignment choices.
Segment Label : *Cited Work - with insight highlighting a negative or positive(gap or problem)*

To deal with the difficulties in parse-to-parse matching , Wu (1997) utilizes inversion transduction grammar (ITG) for bilingual parsing. Bilingual parsing approach looks upon the parsing and alignment as a single procedure which simultaneously encodes both the parsing and transferring information. It is, however, difficult to write a broad coverage 'bilingual grammar' for bilingual parsing.
Segment Label : *Cited Work - with insight highlighting a negative or positive(gap or problem)*

Table 6.4: *Related Work* segmented by the system with segment labels

Bod(2007) reports that the unsupervised STSG-based translation model performs much better than the supervised one. The motivation behind all these works is to exploit linguistically syntactic structure features to model the translation process. However, most of these fail to utilize non-syntactic phrases well that are proven useful in the phrase-based methods(Koehn et al. ,2003).

Table 6.5: Example sentence with a multi-co-reference highlighted in blue. System treats this a one segment where as a human would segment after the first sentence.

type 3, i.e. citation in parenthesis used when evidence is provided when in fact it is of type 1. Particular co-references that were missed were *it* or *they*, or when a system name is referenced rather than the paper, e.g. *Moses*. These seemed to be more frequently missed by the annotators in (Schäfer et al., 2012).

Of the remaining segment errors, ten were caused when a reference to multiple works occurred, e.g. *these approaches, the methods above*. With the current segmentation rules, this type of sentence, because it has a multi-co-reference, will attach to the previous sentence when it has a citation. Human evaluation would separate this type of sentences into their own segment (Example Table 6.5). The human segmented this after the first sentence but the system does not. The remaining segment errors occurred due to problems in discourse markers not present in the list used.

Out of the segments the system got right, only 25 were labelled differently to human labelling. The majority of these were caused by incorrect author intention labels. However, some observed patterns could provide improvements. For example, when an author indicates that they use or are similar to other work, they often follow this by explaining how their work differs. These segments were labelled *Author's Work - based on/uses or similar to other work* or sometimes if the author claims novelty in how they are different the segment would be labelled *Author's Work - highlights novelty*. However, the human labelling chose *Author's work - author's work differs to other work* in both these cases (example Table 6.6). In terms of generating feedback, it is more important to capture the context and hence, preferable that the human applied label is applied. This could be captured by integrating more information from the intention label into the

Geffet and Dagan(2005) proposed an extension to the distributional hypothesis to discover entailment relation between words. They model the context of a word using its syntactic features and compare the contexts of two words for strict inclusion to infer lexical entailment. In principle, their work is the most similar to ours. Their method however differs in that it is limited to lexical entailment and they show its effectiveness for nouns.
Segment Label : *Author's Work - based on/uses or similar to other work*

Similar to our work , Hildebrand et al.(2005) also use information retrieval method for translation model adaptation. They select sentences similar to the test set from available in-of-domain and out-of-domain training data to form an adapted translation model. Different from their work, our method further use the small adapted data to optimize the distribution of the whole training data. It takes the full advantage of larger data and adapted data. In addition, we also propose an online translation model optimization method, which make it possible to select adapted translation model for each individual sentence.
Segment Label : *Author's Work - highlights novelty*

Table 6.6: Example segments labelled author uses/similar to other work but then says what is different or author highlights novelty where human labelled differently.

rule system or a new segment label of *Author's work - similar or uses but differs* to capture all aspects of the segment and generate more accurate feedback.

Overall the system performance is good both for creating segments and applying labels. The errors found, however, do underline the importance of being able to identify co-reference and citation types adequately. The actual feedback generated is assessed in the user study described in the next section.

6.4 *LitCrit* - User Evaluation Study

The goal of the user study is to assess if **LitCrit** has a positive impact on student responses, helping them to identify aspects that they may previously have missed. Differences were assessed by comparing ratings between the results in Chapter 3 and this study. Based on the differences found between experts and students in the first study, it is expected that using **LitCrit** the students will now be much closer in alignment to experts in their ratings and less likely to miss that context is not present.

6.4.1 Participants

All PG students from the first study were invited to take part by email, but only nine students agreed to take part. There are, however, several limitations to using the same participants. Firstly, although time has lapsed between the study in Chapter 3 and the one carried out in this chapter, a time-lapse of nine months, there is still a learning effect. This comes both from the participants having already seen the *Related Work* sections, but also their level of skill as the PG students may have gained more writing and peer-review experience, which could potentially influence their knowledge and responses during this study.

6.4.2 Materials and Task

The *Related Works* used are the same as described in the first study (Section 3.3.4.1). Questions asked during the study can be seen in Appendix B. Tasks are performed online. First, the students are presented with a page that explains the tasks and how the **LitCrit** tool works. Then, the participant is presented with

the *Title, Abstract and Introduction* to read. When they click *Next*, this opens the *Related Work* in **LitCrit** within the survey, initially displaying the *Related Work* with no highlights. Students can toggle between the highlighted *Related Work* and the original text. Following this, students are asked to carry out ratings on a five-point Likert scale as per the first user study (Figure 3.3). A free-text comment box that is labelled *Please feel free to write any comments* is given at the end of each *Related Work*. The *Related Work* presentation is randomised for each student.

Following the *Related Work* assessment, each student is asked a series of questions designed to evaluate the sentence labelling, feedback comments and whether students think using **LitCrit** would be useful in writing a *Related Work*. These were all evaluated on a seven-point Likert scale - Strongly Disagree to Strongly Agree and the design of questions was based on (Brooke et al., 1996). Questions asked can be seen in Figures 6.6 and 6.7 as well as Appendix B. There is a potential bias to be aware of in that our participants have already been exposed to the research of the author through the first study, this might prejudice them to be more favourable towards the system than they would be with no prior knowledge. We discuss this more in Section 6.6 in describing limitations of this study.

6.4.3 Evaluation Method

The Wilcoxon signed paired test is used to determine if ratings changed significantly for students between the studies and Mann Whitney U to test between the students in this study and experts in the first study. Medians and inter-quartile range (IQR) are reported. Again we report effect size with *Vargha and Delaney's A* (Vargha and Delaney, 2000).

6.5 Results

For equivalent comparison, the findings in User Study 1 are recalculated for the 9 students compared to the original 12. All significant results remain.

6.5.1 Rating Comparison

Median ratings and IQR can be seen in Table 6.7. In addition to results from this study (Student 2), included are expert ratings from the first study and ratings for the 9 student participants in the first study (Student 1).

As seen in the previous study there is no significant differences between the average ratings across all documents. We observe that student and expert groups seem much more in alignment when just comparing the scores for each document (Table 6.7) . Particularly observations in Table 6.7 show context and citation evaluation ratings appear to differ for students after using **LitCrit**. Context ratings look lower in this study for the students and citation evaluation looks lower for *Related Works* D and F and higher for *Related Work* A. Testing rating tendencies, i.e. are students in this study more likely to rate items higher or lower than they did in the first study, only context is significantly lower for students in this study, $p < 0.05$ (V = 1483, p-value = 0.0122). Unlike the first study, there are now no significant differences, using **LitCrit**, in any of the rating tendencies between students and experts. Students in the first study struggled to notice context and any mention to the author's work was missing in *Related Work* F. After using **LitCrit**, students have a significantly lower regard for *Related Work* F's quality, context and support than previously, $p < 0.05$ (Quality W = 9.5, p-value = 0.0275, Context , W = 9.5, p-value = 0.002, Support W = 16.5, p-value = 0.009). All of these results have a large effect size (cf. effect size table for VDA in Chapter 3, Table 3.5) comparing the student groups before and after using *LitCrit* for Paper F Quality, VDA - 0.12, Context, VDA -0.12, Support, VDA- 0.2. Students no longer differ significantly in any ratings of *Related Work* F or any other *Related Work* compared to experts. However, *Related Work* B does remain higher in median quality rating. This is likely due to what was found in Section 3.7.2, unlike experts, PG students do not realise that many of the citations are not relevant thus giving it a higher rating, although this is not significantly higher.

6.5.2 Discussion - LitCrit Results

Does highlighting author intentions and giving feedback change the ratings given by the students?

ID	Group	Quality	Context	Cit Eval	Support	Detail
A	Expert	2 (0)	1 (1)	2 (1)	4 (2)	2 (0)
	Student(1)	2 (1)	2 (1)	2 (1)	4 (1)	2 (0)
	Student(2)	2 (1)	1 (1)	3 (1)	5(1)	3 (1)
B	Expert	2 (1.5)	2 (1)	3 (1.5)	4 (2)	3 (1)
	Student(1)	3 (2)	3 (1)	4 (1)	4 (1)	3 (0)
	Student(2)	3 (0)	2 (0)	4 (2)	4 (0)	3 (1)
C	Expert	4 (1.5)	4 (1.5)	4 (2)	4 (1)	3 (0.5)
	Student(1)	4 (1)	4 (1)	5 (1)	4 (1)	3 (0)
	Student(2)	5 (1)	4 (1)	5 (1)	4 (1)	3 (0)
D	Expert	2 (1)	2 (2)	3 (2.5)	4 (2)	2 (1)
	Student(1)	2 (1)	3 (1)	4 (1)	4 (1)	2 (1)
	Student(2)	2 (1)	2 (1)	2(1)	4 (0)	2 (0)
E	Expert	4 (2)	4 (0.5)	4 (0.5)	4 (1)	3 (0)
	Student(1)	4 (1)	4 (1)	3 (2)	4 (1)	3 (0)
	Student(2)	4 (1)	4 (1)	3 (2)	4 (1)	2 (1)
F	Expert	2 (1)	1 (1)	3 (1.5)	4 (1.5)	3 (.5)
	Student(1)	4 (1)	2 (2)	4 (3)	5 (1)	3 (0)
	Student(2)	2* (1)	1* (0)	3 (2)	3* (0)	3 (2)
G	Expert	3 (1)	3 (2)	3 (1.5)	4 (2)	3 (1)
	Student(1)	3 (0)	3 (1)	2 (1)	4 (1)	2 (1)
	Student(2)	3 (1)	3 (1)	2 (0.5)	4 (1)	2 (1)

Table 6.7: Agreement on Ratings for *Related Works* A,B,C,D,E,F by Expert - results from the first study in Section 3.2, Student(1)- results from the first study in Section 3.2 recalculated for the 9 students in this study, Student(2) - results from this study. Medians for scores are reported (Likert Rating of 1 being the lowest and 5 being the highest) with Inter-Quartile Range (IQR) in brackets, significance ($p < 0.05$) between students in the different studies is denoted by *. No significant differences are found between Student 2 and the Expert group

In the first study, the students were observed to be less likely to notice that context was missing in *Related Work* through meaningful comparison of cited works to the author's work or that the author had failed to mention their own work. Evidence is found that **LitCrit** influences student context ratings, showing significant differences between the studies in overall context rating tendencies. Additionally, students using **LitCrit** show no significant difference in ratings with experts, unlike the first study.

Related Work F in particular caused problems for students in recognising context or the author's failure to mention their own work. Using **LitCrit**, significant changes in both quality and context for *Related Work* F are observed. The highlighting of the narrative with author intentions has helped the students look beyond the superficial aspects of the writing, drawing attention to missing aspects.

While **LitCrit** draws attention to citation evaluation, it was thought this might not influence quality ratings because students view citation evaluation purpose differently to experts. Differences are observed in the median ratings of *Related Works* A, D, F between the studies but these are not significant between the two student groups. However, referring back to Table 3.1, which describes the assessment of criteria of the *Related Work*, *Related Work* A does include critical evaluation. It is reasonable to conclude that highlighting this within **LitCrit** might have led to students raising ratings of citation evaluation of *Related Work* A. Conversely, as *Related Work* D contains no critical evaluation, it is reasonable to assume **LitCrit** highlighting may have caused students to lower ratings for *Related Work* D. There still remains an issue with citation evaluation, likely stemming from what was found in the first study, i.e. students view its purpose as listing limits and merits but do not strongly link it with quality. It is possible that attention can be drawn to this by how feedback comments are phrased or in some pre-instructional information about what makes a good *Related Work*. However, there does seem a wider need to address this from a pedagogical point of view.

6.5.3 User Perception of LitCrit

6.5.3.1 Author Intention Sentence Labels and Highlighting

Sentences highlighted with author intention were thought to be helpful and drew attention to aspects (Figure 6.6). Free-text comments support this and show the

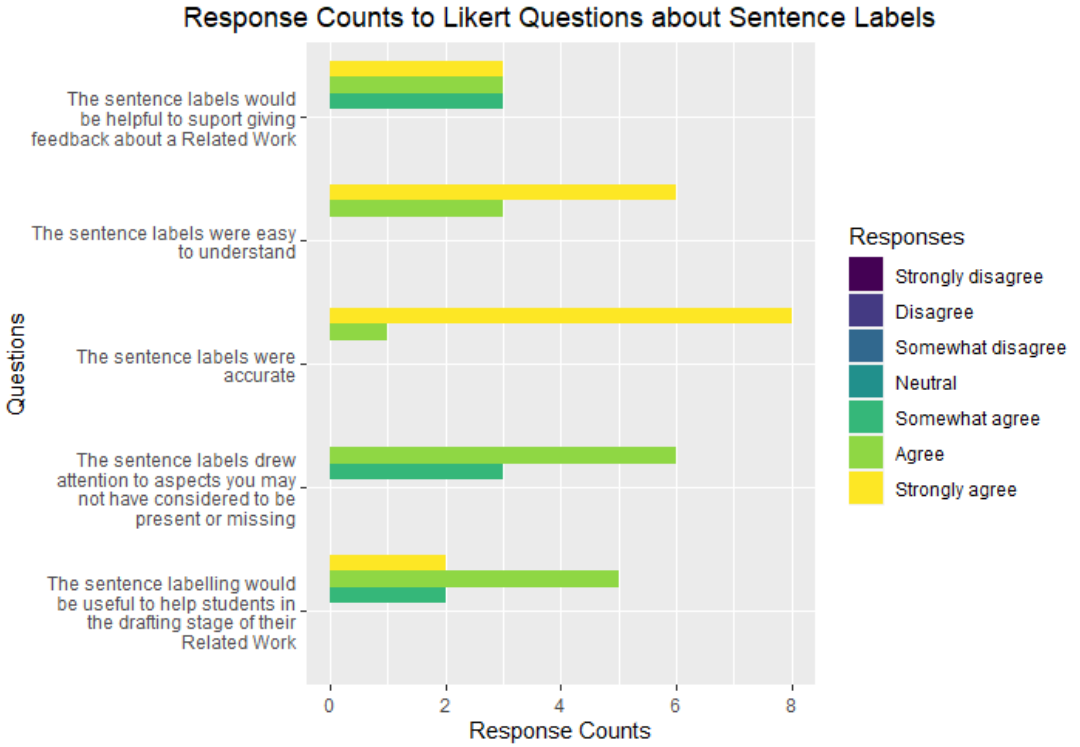


Figure 6.6: LitCrit user evaluation with the questions asked of the participants about the author intention sentence labels and their responses, rated on a 7 point Likert scale Strongly Disagree to Strongly Agree.

highlighting helped students challenge their idea about what the content represented and how it linked to the author intention label:

- *the labels were always helpful in highlighting phenomena even in the rare cases where I then mentally decided a different label might have been applicable [S1]*
- *It makes you consider if your first thoughts about the text are correct, or whether you should challenge your view or not [S6]*
- *The highlights were really useful for getting a first impression of the structure and balance of the data [S9]*

- *Tags definitely helped break this one down [B] due to the high amount of background information given [S6]*

There were some free-text comments from participants about sentence label accuracy. This shows that, despite the manual review for accuracy, disagreement may still arise as arguably sentences could take multiple author intentions labels:

- *There are some cases where arguably a different sentence label could have applied than the one that was chosen, but I would not say that any of the labels were *in*accurate, just that in some cases there might be more than one label that could plausibly be applied[S1]*
- *LitCrit missed the implied criticism of "Their model needed large-scale corpora to estimate the probabilities and to prevent data sparseness"[S9]*

There was a very strong positive response to *The sentence labels would be helpful to support giving feedback about a Related Work*. This is an aspect we had not considered originally when designing **LitCrit**. However, it seems that in addition to supporting writing, the approach is also helpful for novices trying to read or give feedback. This could prove a different but useful angle to pursue with **LitCrit** in the future.

6.5.4 LitCrit Feedback Comments Box

User evaluation for the **LitCrit** generated feedback (located in the comments box, Figure 6.2) can be seen in Figure 6.7.

Overall students seemed to have found this useful. However, there was a lower rating for the last point about highlighting or bringing to attention aspects that may be missing. This may be due to the feedback being somewhat repetitive, given that the author intentions were already highlighted. This is a useful point for future consideration to make sure the feedback adds value and does not just re-iterate the information the author intention labels provide:

- *I found the comments box to be somewhat helpful at first, but more so just reiterating what seemed clear from the labels themselves....The comments box seems more useful for novices in that respect[S4]*
- *The sentence labels and suggestion box are complementary and work very well together[S1]*

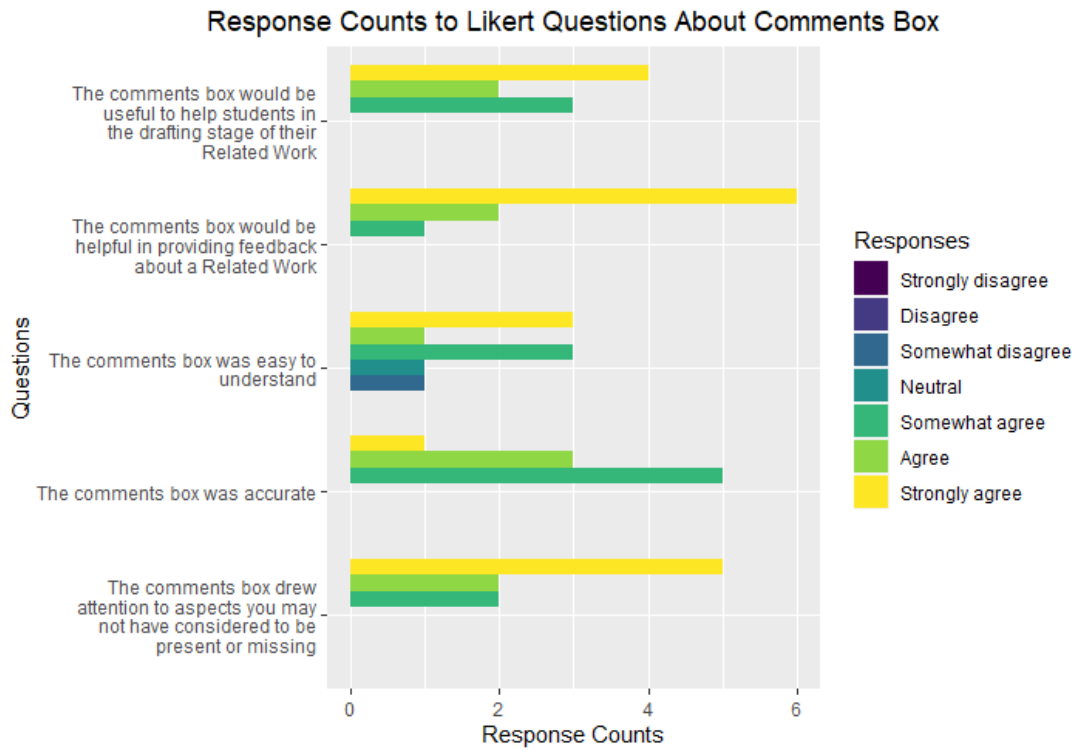


Figure 6.7: LitCrit user evaluation with the questions asked of the participants about the feedback comment box and their responses, rated on a 7 point Likert scale Strongly Disagree to Strongly Agree.

Overall, highlighting the narrative using author intentions was well-received, helping the students to identify characteristics present and challenge them to think deeper about the content. General comments indicated that overall students thought that **LitCrit** would be a useful tool: *I genuinely think this would be a very useful tool indeed![S1]*, *It would be extremely useful.[s4]*

6.6 Conclusions and Limitations

Using the findings developed earlier in this thesis, we show in this chapter that focusing on content expectations through author intentions and visualising these in the *Related Work* narrative helps PG students identify aspects they previously missed. There is a significant change in context and quality ratings when PG students use **LitCrit** compared to when they did not. Unlike in the first study, no ratings for students now show any significant differences to that of experts. Thus, providing evidence that using **LitCrit** influences the PG students thinking about

Related Work, bringing ratings more in line with those of experts. However, while not significant, we do still observe some differences in ratings for critical evaluation. Drawing attention to critical evaluation labels in the *Related Work* appears to have influenced the PG students to increase or decrease their scores according to labels present. From the results in Section 3.7.2, PG students seem to have a different understanding of citation evaluation purpose than experts, focusing on limits and merits rather than how this evaluation needs to be put in context to the discussion. This deeper mis-understanding may lead to these higher and lower ratings than experts and may require pedagogical intervention or more instructional information within **LitCrit** to further support writing.

Overall, **LitCrit** was received positively in the user evaluation, and PG students thought it would be of assistance in writing a *Related Work*.

6.6.1 Limitations and Future Work

Although the outcome from our study is positive the small sample size limits the generalisability of these results (cf. Section 3.8.1). Additionally, the use of the same participants will introduce some bias and a possible learning effect. Although nine months have passed, the participants had already seen the *Related Works* used. Additionally, the participants may have developed or improved their knowledge during that time regarding *Related Work* writing. With **LitCrit** now being in place, future work could be done, which could reduce the potential learning effect using more participants and *Related Works*. In addition, an evaluation task of the **LitCrit** system without the peer-review task could provide further insight into the usability of the system.

The evaluation we undertook was based on manually corrected labels and was not fully automated. This correction could lead to optimistic results, whereas when full automation is introduced the results would be less favourable, as users noticed mis-labelling, or did not agree with **LitCrit**. However, PG students did highlight issues with sentence label accuracy. Whilst multi-labels could be found in a sentence, PG students also pointed out that disagreeing with a label was not necessarily a bad thing, as this made them challenge their thinking. Additionally, we did acknowledge that our approach to providing feedback was fairly simplistic, being a template-based approach. Some students commented that

the feedback comments were iterative and perhaps not as valuable as the labels themselves. The aspect of providing the labelling and feedback needs to be evaluated further in future work to see how much value it provides and if it could be improved. We explore this more in Section 9.2.1 where we consider future work that could be done to improve **LitCrit**.

At this time, **LitCrit** does not focus on language clarity, grammar or presentation, all critical aspects of writing. These may impact **LitCrit**'s ability to identify author intentions when it is fully automated, and thus may need to be integrated in the future.

Chapter 7

Discipline Independence of the Author Intention Framework

7.1 Introduction

This chapter explores how well the classifier developed in Chapter 5 performs on automated recognition of the author intention labels in the discipline of Computer Graphics. The experiments undertaken so far in this thesis have all focused on the discipline of Computational Linguistics (CL). The earlier background discussion drew attention to the challenges of working across disciplines due to the variation in language and presentation of scientific arguments. These differences mean that any framework and its features may not perform well within another discipline and adaptations are often needed.

7.2 Adapting a Model to a New Domain

Often, we come across problems when a model trained for one domain may poorly generalise to a new domain. Additionally, methods of using supervised learning may be problematic when we do not have sufficient labelled data for the new task or domain of interest. Transfer learning is an approach to solve these problems, leveraging the already existing labelled data of some related task or domain. This allows us to take a pre-trained model for a task or specific domain and use it for another task or within a different domain.

Pre-training of models often includes an auxiliary task that allows the model to learn. One common method in NLP tasks is to use word vectors (Mikolov et al., 2013) that map word identities to a continuous representation where similar words map to similar vectors. A common approach to learning these vectors is to favour co-occurring words to be positioned nearby in the continuous space (Mikolov et al., 2013). Recently, it has become more common to pre-train an entire model on a task with a large existing data resource. This pre-training causes the model to develop the ability to generalise and build knowledge that can then be transferred to downstream tasks. This approach has led to state-of-the-art results in many of the most common NLP benchmarks (Devlin et al., 2018; Liu et al., 2019b; Lan et al., 2019; Peng et al., 2019). In addition to its ability to generate state-of-the-art results, these types of methods widely appeal, as they reduce the need for expensive and time-consuming pre-labelled data. This pre-training method is a natural fit for neural networks, which have been shown to exhibit remarkable scalability, i.e. it is often possible to achieve better performance simply by training a larger model on a larger data-set.

Our existing model is based on hand-crafted features, and such features tend to be bounded by a specific domain (Hussein et al., 2019). Whilst our feature-engineering approach may not benefit from the transfer learning described above, our reasoning for this approach was that we were more interested in explaining relationships between features and outcomes of our models, and learning any pedagogical aspects that could better support writing feedback. Our approach therefore in this chapter is simplistic in that we consider how well the existing model performs in this new domain, and, through our understanding of the features the model uses, we remove some very domain specific features to improve the model in this new domain. Despite the simplistic approach, this is valuable, as it also allows us to understand if and where these features exist in this new discipline, and consider how this may relate in general to providing writing support for a *Related Work*.

7.3 Computer Graphics Data

7.3.1 Description of Data

The data set of articles used in this chapter is from the field of Computer Graphics (CG). This field includes work that considers the use of visual content in a computer and covers anything related to the generation or manipulation of this type of content. The papers used are described and used in several previous experiments (Fisas et al., 2015, 2016). These papers were the basis of developing the *ArguminSci* schema, described earlier in Section 2.2.1.4 and compared to the intentions in this thesis in Section 4.5. The data set of articles used in this chapter was requested and supplied by the authors of the paper (Fisas et al., 2015). This original collection of papers, 40 in total, were selected at random by (Fisas et al., 2015) from a larger collection, designed to be representative of articles in Computer Graphics. The papers are classed into four subject areas: Skinning, Motion Capture, Fluid Simulation and Cloth Simulation. The final articles we use in this chapter is a subset of the original data set, 37 papers. Two papers were removed as they did not have *Related Work* sections and the third was removed as the *Related Work* contained a theoretical background description with equations rather than a discussion on *Related Work*.

7.4 Annotating the Data

These papers had to be annotated not only to apply the author intention labels but also for co-referencing. The annotator was the author of this thesis. Previously mentioned is the challenge of having annotators with too much domain knowledge as they can often infer knowledge when no linguistic clues are available in the text. The author of this thesis does have an elementary knowledge of computer graphics but not to the level of techniques or methods detailed in the paper set.

7.4.1 Annotating Co-references

The original annotation for co-referencing of the CL data was done according to (Schäfer et al., 2012). The processing step of how this is converted to co-

reference annotation markup for this work has already been discussed in Chapter 4. This thesis is not about developing a co-reference tool and therefore the annotation here is done mainly as a manual process although some pre-processing is used to support the annotation process.

Pre-processing of Data

The data set of 37 papers is provided by (Fisas et al., 2015) as XML data, pre-processed into sentences. We extract the sentences along with the document ID and assign a sequential sentence ID. The sentence is passed through the citation parser previously described in Section 5.3.4 to identify citations of the three types: CIT1, CIT2, CIT3. This replaces the citations with the CIT[1-3] placeholder. This fields of *Document ID, Sentence ID, Sentence, Sentence parsed with citations marked*, are then saved into Excel.

A list of the co-reference annotations which contain possessive pronouns, e.g. *their experiment, their model, our model* is extracted from annotations in the (Schäfer et al., 2012) data set. This is used to parse all sentences, and any matches are highlighted in bold and used as a prompt for the annotator to help identify co-references.

Annotation of Co-references

The sentences from the *Related Work* are presented in an Excel file. Each row represents a sentence with fields corresponding to document ID, sentence ID, the original sentence, a column with the original sentence parsed for citation markers, e.g. Smith et al. is replaced with CIT1, and any potential co-reference highlighted, and a final column where the annotator could mark the sentence ID that first introduced any citation being co-referenced. The annotator replaced the original text, e.g. *this paper, our work, their work* with the appropriate co-reference marker, described in Section 5.3.3. This was repeated three times to minimise any errors.

7.4.2 Annotating for Author Intention Labels

Annotation for author intention labels is carried out using the same guidelines described in Section 4.7.3 and follows a similar process as annotating the co-references.

The *Related Work* sections were presented in an Excel file. Fields 1-4 of each row representing a sentence as noted earlier. Field 5 is for the annotator to enter a label from the pre-populated list provided. Field 6 is for comments. As before the annotation was repeated three times to minimise any errors.

7.5 Discipline Differences and Label Distribution

The label distributions presented in Table 7.1 show the percentages of each label for each discipline. We use χ^2 test of independence to examine the relation between the disciplines and author intention labels. We also measure the effect size using Cramér's V (Cramér, 1946). The χ^2 test was significant ($p < 0.01$), $\chi^2 = 132.90$, df=9, p-value <2.2e-16 and Cramér's V = 0.23, which can be interpreted as strong effect. Thus, this is an indication that the disciplines have significantly different relations of author intention labels with a strong effect.

Label distribution (Table 7.1) and our observations during annotation give us insight into the differences between the disciplines. Background labels are more prevalent in this new discipline, as are evaluative sentences for the background. This reflects the different style of *Related Work* in this discipline. Authors talk about techniques and models in general, using citation for evidence rather than single descriptive cited works sentences. Authors appear to provide more critical evaluation about these techniques saying what is good and bad but also making clear what their contribution is. Many of the *Related Works* have a sub-section entitled *Author Contribution*. This was not found in any of the CL papers. It seems there is a stronger expectation of discussing the author contribution in the CG discipline, resulting in many more author contribution (A-Gap) sentences. However, there are fewer citation description (CW-DESC) sentences. There are also fewer sentences that compare the authors work to a cited work (A-CW). This is probably a reflection that the authors discuss more citations in general as techniques or methods and then state clearly how their work differs, rather than

using one sentence to compare their own work to a single citation.

Sentence Label	ACL Papers %	Comp Graphics Papers %
BG-NE	14.60	17.00
BG-EP	9.70	14.00
BG(-+)	5.00	12.00
CW-DESC	40.00	24.40
CW(-+)	7.50	7.80
A-USE	3.40	4.60
A-DESC	6.00	6.50
A-CW	8.60	5.00
A-GAP	3.40	6.50
TXT	1.10	1.20

Table 7.1: Label class percentage distribution for ACL papers used in Chapter 5 and the label distribution percentage for the Computer Graphics papers described in this chapter. Percentages are used as the number of papers differs.

7.6 Experiments

7.6.1 Methods

The four experiments described next are undertaken to understand how well the existing model classifies the new unseen data from Computer Graphics *Related Works*. All sentences from each article are processed to produce the feature of sentence vectors described in Section 5.3 and 5.9 for input to the classifier experiments. The exact same features and model classifier configuration is used as described in Chapter 5:

- All models are trained using LibSVM (Chang and Lin, 2011) with a linear kernel and default settings with 10-fold cross validation.
- Significance is tested using the corrected t-test (Nadeau and Bengio, 1999) $p < 0.01$.
- All features used are as described in Section 5.3 and 5.9. Where any features are omitted, this is highlighted in each experiment description.

7.6.2 Experiment Description

Experiment 1 In this experiment, the CG papers are used as a test-set for the best performing model of Section 5.9.2. This is the very last model described, which reaches an accuracy of 76.34% and uses the new subject guess and the label suggestion feature.

Reported are the results, overall accuracy and individual F1 scores for labels for this classifier model on the original CL data (cf. Section 5.9.2) and using CG data as a test-set.

Experiment 2 Previous discussion in this thesis has highlighted the known issues of language variation between disciplines, and in this chapter, the annotation task also highlighted presentation differences better CG and CL, with CG being more explicit about their own contribution. The location of this discussion often differs coming as a sub-section at the end of the *Related Work*. In order to try to remove some of the discipline specific features, the original classifier is rebuilt removing the following features: n-grams, dependencies, previous label and label suggester. The CG data is then used as a test-set for this model.

Reported are the results, overall accuracy and individual F1 scores for labels for this classifier model on the original CL data and using CG data as a test-set.

Experiment 3 The Computer Graphics papers are run as a classification experiment on their own. All features are used except the CL specific label suggester feature.

Reported are the results, overall accuracy and individual F1 scores for labels for this classifier model.

Experiment 4 Label distribution differs between the two paper sets and it is possible the sparseness in one may be offset by including more examples from the other discipline. This experiment tests if including samples from another domain is helpful in improving the accuracy of the classifier. Both data sets are combined, and the classifier model rebuilt using all features except the domain specific one of n-grams, dependencies and the label suggestion feature as this is domain specific to CL.

Reported are the results, overall accuracy and individual F1 scores for labels for this classifier model.

7.7 Results and Discussion

This section describes the results of all four of the experiments presenting and comparing them. Results are presented in Table 7.2 and 7.3.

Features	BG(-+)	BG-NE	BG-EP	CW(-+)	CW-DESC	A-USE	A-DESC	TXT	A-CW	A-GAP
Experiment 1										
OrigCL	44	78	81	68	90	64	65	71	54	34
CG-Test	28	23	72	30	52	21	6	0	4	3
Experiment 2										
OrigCL(adapted)	37	72	79	60	87	60	47	69	64	25
CG-Test	35	54	**89**	44	67	23	0	0	48	03
Experiment 3										
OrigCL	44	78	81	68	90	64	65	71	54	34
CGOnly	**48**	58	**91**	51	82	56	53	18	48	**43**
Experiment 4										
OrigCL	44	78	81	68	90	64	65	71	54	34
Combined	**51**	74	**82**	63	90	58	54	41	**62**	**46**

Table 7.2: F1-Measures (%) for labels in Experiments 1 - 4

7.7.1 Experiment 1

Using the best performing model for the CL data and CG as a test set, the results are poor, only reaching an overall accuracy of 40.17% (Table 7.3). Only two labels generate F1 scores that would be reasonable BG-EP at 72% and CW-DESC at 52%. These results would not be reliable enough to generate feedback.

7.7.2 Experiment 2

Accuracy improves by almost 15% when some of the very discipline specific features of language are removed, e.g. n-grams, dependencies, previous label. However, the majority of the labels would still not be reliable enough for feedback, particularly those of A-DESC and A-GAP(0 and 3%). These disciplines differ in the way they discuss their own work in a *Related Work* section and this is reflected in the results with these labels being harder to classify.

Features	Precision%	Recall%	MicroF1%	MacroF1%	Accuracy%
Experiment 1					
OrigCL	75	76	75	65	76.34
CG-Test	51	40	36	28	40.17
Experiment 2					
OrigCL	70	72	72	60	71.72
CG-Test	53	54	50	36	54.12
Experiment 3					
CGOnly	64	65	54	55	64.73
Experiment 4					
Combined	74	74	74	62	74.23

Table 7.3: Precision, recall, overall F1 and accuracy (%) score for all experiments with comparisons to the original results in Section 5.9

7.7.3 Experiment 3

These results are much more encouraging. Using the CG data on its own, although with an overall 10% lower accuracy than the best performing model on CL data, shows improvement in individual F1 scores for labels, particularly A-GAP and BG(+-). The label distribution is different in the CG discipline having more labels in BG(+-), and A-GAP are most likely what contribute to the increased performance seen in F1 scores for these labels. There does, however, remain a problem with TXT. The lexicon category (TXT_NOUN) for this is particularly good at picking up words that indicate a TXT sentence in CL. The majority of CG sentences that fall into this category do not have any words that match the lexicon. This indicates that this type of sentence is very discipline specific.

7.7.4 Experiment 4

The combined data set results are very close to the best performing CL model, but the increase in A-GAP and BG(+-) label is particularly encouraging with this model producing the best F1 scores for 4 out of the 10 labels.

7.8 Discussion and Conclusions

This chapter considers how discipline independent the model previously built is, applying it within a new domain of Computer Graphics. Overall the model does not perform well in the new domain and there is a need to choose features that rely less on linguistic variation or difference between the structure of the text. This was demonstrated in Experiment 2 with increased performance when features such as n-grams and previous label are removed. Given what is known about variation across disciplines, these results are not surprising. Our work has focused on building features that align with the fine-grained aspects of our author intentions, specialising on *Related Works*. This problem seems to be in distinguishing features that are more domain specific compared to those that are more generic to *Related Work* and an understanding of the differences that occur in sections between the disciplines is needed. For example, the CG domain has a different presentation structure of intentions to the CL domain with larger subsections at the end discussing the author's work. This difference resulted in the *previous label* feature being detrimental to the performance, shown through Experiment 1 and 2, but it is better to include this feature in Experiment 3 when only the CG data is used to build the model.

Combining data from the two domains, in Experiment 4, to build the classifier overcame some of the problems of label sparseness and improved F1 scores for individual labels, such as A-GAP. The idea of being able to sample data or provide more samples to boost problematic labels could be useful in future. Comparing the author intention model developed in this thesis to other models (cf. Section 4.5) there are some existing models with labels that are similar to ours. It could prove advantageous to explore how to use these to expand the training set and avoid expensive annotations.

Other tools described in this thesis such as *Research Writer Tool* and *Criterion* (cf. Section 2.3) claim to be discipline independent. However, these are built on much larger training sets than available in this work. This experiment does support the notion of discipline independence if training data sets can be grown to sufficient size and are trained on all disciplines. Overall the results are encouraging, but they underline the need to have enough training data to overcome sparseness and discipline differences and that training a model for one discipline means it will not necessarily work *out of the box* on an unseen discipline, thought

needs to be given to the feature set.

Chapter 8

Predicting Quality with Author Intentions

8.1 Introduction

This chapter uses the author intention labels to predict the quality of a *Related Work* showing the intentions serve as a proxy for the content that is expected to be present. The corpus of *Related Works* used for annotation are rated as Good, Poor or Fair in an assessment study and the relationship of the author intentions to the quality ratings is analysed. The work in this chapter is published in (Casey et al., 2019a).

8.2 Author Intention as a Proxy for Quality

Argumentative elements identified through discourse analysis have been successfully applied to automatic determination of student essay scores (Burstein et al., 2004; Song et al., 2014). Ong et al. (2014) develop a rule-based argument ontology to parse sentences within texts. Despite its rather simplistic approach, which is mainly based on discourse connectives and a small corpus of 52 papers, they are able to demonstrate that these elements were related to higher scores. The Criterion writing tool (cf. Section 2.3) incorporated writing intentions into their score predictor and showed this contributes to predicting an essay score. However, in this work, the focus is not on predicting a score but an indication of

quality. In this work, we have used peer-review to assess what content experts look for in a *Related Work* and mapped this into a model of author intention labels. In theory, using these intention labels should, therefore, be a good predictor of quality and differences in the occurrence of labels between good and poor *Related Works* should be visible. This prediction of quality should provide evidence that the mapping of author intention labels is indeed a good representation of content expected by experts.

8.2.1 Problems with Judging Quality

Automating judgement of quality in research, though, is challenging as it requires knowledge. Bridges (2009) describes this judgement of research quality as a connoisseur-ship which draws on one's own knowledge and experience of the field. This, in turn, not only allows one to comment on specific features but also gives one the ability to appreciate the overall composition of the text. Evidence from our first user study (Section 3.7.2) supports this. Unlike the experts, the PG students failed to notice that most of the cited work in *Related Work* B was not relevant to the author's work, given the *Abstract* and *Introduction*. It is not an aspect that our author intention labels could detect either.

The experiment we undertake to rate quality is based on peer-review. Peer-review is generally accepted as the gold standard of assessing quality but it is not without issue, such as bias with regard to an individual's attitude to the material in the paper (Walker and da Silva, 2014), but also bias introduced because of our convenience approach (Kelly, 2009, Ch. 7, p. 67) to recruit the participants. This approach may influence the participants' responses about quality, as some subjects are familiar with the thesis author's topic. However, the majority of the participants were experienced researchers and we believe that this would have been forefront in their judgements of quality.

Metrics, such as citations and download counts, could have also been considered as indicators of quality. However, these have known issues, such as dependence in the discipline size, time taken to accumulate. Authors and research teams have been known to carry out unnecessary self-citations to increase their own citations (Glanzel et al., 2006). Some would argue that these metrics are not directly related to quality but measure impact – after all a citation may be there

to say something negative.

Another problem is that the author intention labels do not take into account the presentation or structure aspects that are expected. Thus, they cannot capture a full picture of the quality. While it is difficult, if not impossible, to try to emulate human judgement in an automated fashion, it is still useful to measure how well the author intentions represent quality. Using the author intentions will not fulfil all the requirements to predict quality, but this work studies what contribution it can make.

8.3 Assessment Study

8.3.1 Material and Procedures

An experiment was set up to rate the quality of each *Related Work* section from the annotated data set previously described in our annotation study (cf. Section 4.6).

Materials were all presented on-line. Participants logged in and were presented with a list of *Related Works* assigned to them. Clicking on an individual assigned *Related Work* presented the participant with the *Title, Abstract* and *Related Work* section, seen in in Figure 8.1. The participant uses a toggle button at the side which toggles between the initial information of *Title, Abstract* and *Related Work* and the assessment questions. Assessment questions, depicted in Figure 8.2 asked the participant to:

1. Rate the overall quality into Poor, Good or Fair
2. Indicate whether enough related work had been cited
3. Assess how well the cited works have been related to the current (author's own) work
4. Indicate how clear the differences between the current work and the author's own was
5. Add any additional comments as to what they thought made this a good or bad paper

In this current assessment, only item 1 (overall quality) is used.

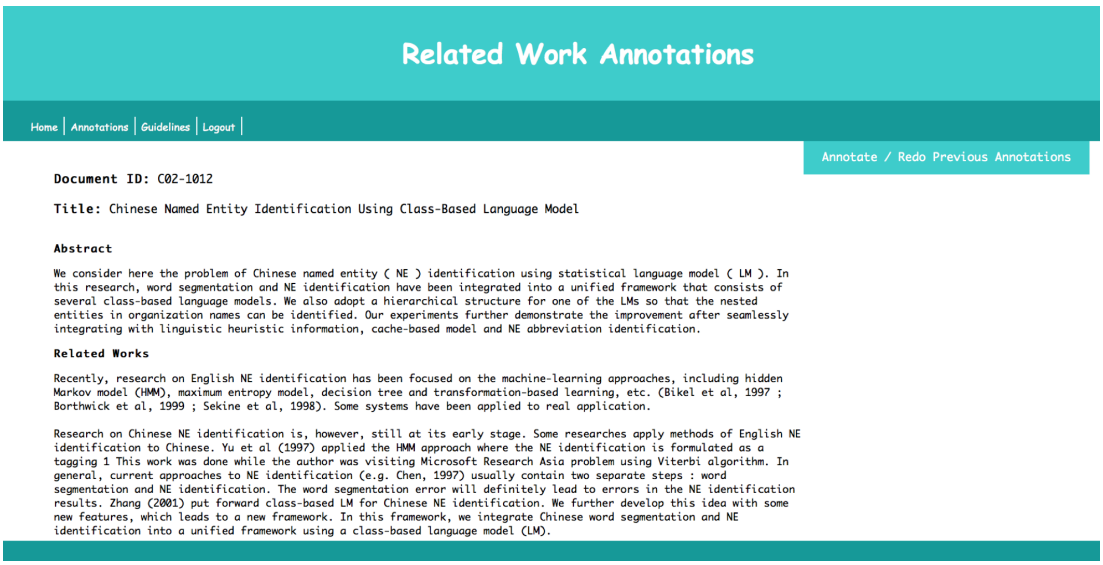


Figure 8.1: Screen shot from the system used to rate the *Related Work*. the screen presented is used to read the *Title, Abstract and Related Work* and the toggle button can be used to see the questions used for ratings.

Guidance given to participants about quality of *Related Work* sections suggested that it was not enough to list previous work, but that authors should demonstrate the relation of cited work to their own work. This guidance also indicated that the *Related Works* were from 6-8 page conference papers so an in-depth explanation of state of the art in the *Related Work* was not expected.

8.3.2 Participants

There were six participants: four experts and two PhD students – all in the Computational Linguistics discipline except one student in Computer Vision. Participants were recruited through the author's academic network at the School of Informatics. This method of convenience sampling (Kelly, 2009, Ch. 7, p. 67) can introduce bias, such as the participants being primed toward the author's work and prejudiced by any previous knowledge of what in this instance, the thesis author views of what quality was.

Ethics approval was obtained for the experiment by the School of Informatics Ethics Committee.

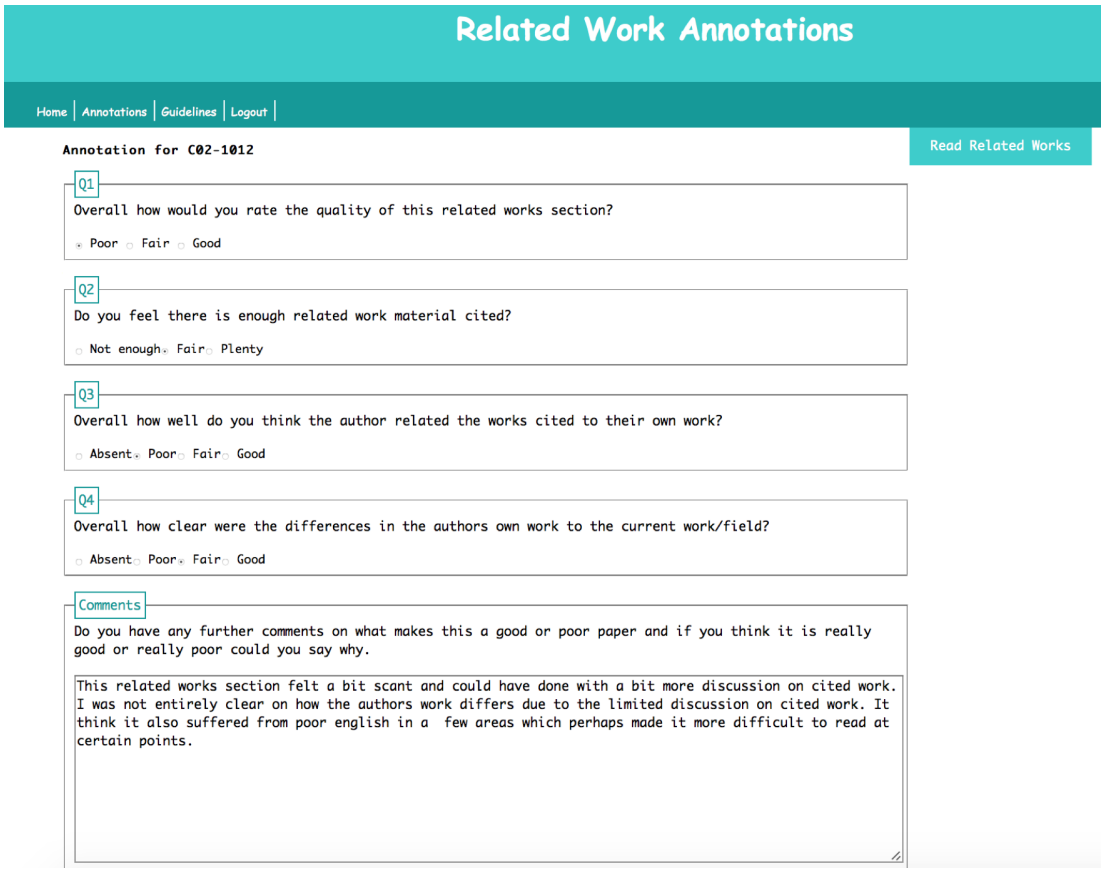


Figure 8.2: Screen shot from the system used to rate the *Related Work*. The screen presented is used to make the rating about the *Related Work* and the toggle button (Read Related Works) can be used to see the *Title, Abstract and Related Work* screen.

Label	Poor	Fair	Good	Significance
BG-EP	1.2 (0.7)	2.0 (2.0)	2.5 (5.1)	* - *
BG-NE	2.2 (10)	3.4 (5.4)	2.0 (4.5)	- * -
BG(-+)	0.8 (1.4)	1.4 (3.7)	1.2 (2.5)	- - -
CW-DESC	8.0 (46)	8.0 (35)	5.6 (21)	- - -
CW(-+)	1.3 (2.0)	2.3 (5.2)	1.3 (3.2)	* - -
A-USE	0.4 (0.3)	0.6 (0.7)	1.0 (1.3)	- - *
A-DESC	0.5 (0.9)	1.5 (2.4)	1.4 (2.7)	* - *
A-CW	0.2 (0.2)	1.2 (1.4)	3.7 (3.7)	* * *
A-GAP	0.1 (0.3)	0.5 (0.5)	1.4 (1.2)	* * *
TXT	0.2 (0.9)	0.2 (0.2)	0.3 (0.3)	- - -

Table 8.1: the table shows the mean occurrence with variance shown in brackets for each sentence labels by rating. Significance is denoted by * in the right of the table, ordered by Poor/Fair, Fair/Good, Poor/Good.

8.4 Assessor Agreement

One assessor rated all items, and the others rated ten each. Assessor agreement considered the differences between the five assessors and the main assessor who looked at all the articles.

Four out of the five assessors were in good agreement with the main assessor; two were in complete agreement, and two agreed on seven out of the ten papers. The other assessor only agreed in four instances, which is likely due to them being a PhD student in another area and having less experience with ACL papers. All disagreements were discussed, and agreement was reached, resulting in 50 double rated papers and 44 done by one assessor only. This resulted in a final data set of 94 *Related Work* with Poor-(36%), Good-(31%) and Fair-(33%).

8.5 Mean Label Occurrence in Rated Sections

Table 8.1 shows the mean number of times a label occurs in each section, grouped by quality rating with variance in brackets. The intuition is that the occurrence of some labels will vary between the different types of ratings. Welch's t-test is

used, correct for unequal variances, to test if differences are significant between the means in the groupings. Each group is tested in order of Poor/Fair, Fair/Good and Poor/Good, where * denotes the test was significant ($p < 0.05$).

The background label with evidence (BG-EP) in Poor sections is found to be significantly different from those that occur in Fair or Good rated sections. There is a significant difference in the number of background statements in Fair rated sections compared to Good sections that provide no evidence (BG-NE). Experts in our study clearly pointed to expectation of context being made clear in *Related Work* and the findings in Table 8.1 support this in terms of significant differences between the mean number of sentences in a Good rated section that describes how the author's work is different to a cited work (A-CW), and how the author's work fills a gap (A-GAP). Additionally, there is a significant difference in the number of sentences that describe an author's work (A-DESC) in Poor rated sections compared to both Fair and Good sections.

8.6 Experiment Methods

Related Work quality is classified into Poor, Fair or Good. Three classifier are trained,: SVM (linear kernel, default settings), Decision Tree (C4.5) and Linear Logistic Regression (LLR) (Chang and Lin, 2011; Quinlan, 1993; Sumner et al., 2005). Feature sets are the annotated labels only. While there are many other features that could be included, the focus here is to understand how well the author intentions relate to quality ratings. 10-fold cross validation is used and a majority classifier is used as the baseline. Classifier precision, recall and accuracy, mean performance over 10 iterations with variance is reported. Also reported is how the label features rank in terms of importance in the best performing classifier. Significant difference testing between the classifiers is done using corrected t-test, ($p < 0.05$) (Nadeau and Bengio, 1999).

8.7 Results

Table 8.2 shows precision, recall and accuracy from all three classifiers and the majority class baseline. All classifiers outperform the baseline significantly. Unsurprisingly, SVM and LLR produce similar results. However, SVM displays

Classifier	Precision%	Recall%	Accuracy%
LibSVM	70 (1)	70 (1)	70 (1.90)
J45	60 (4)	60 (5)	57 (5)
Logistic Regression	70 (2)	70 (2)	70 (2.20)
Majority Baseline	-	36	36

Table 8.2: Classifier performance for each method used showing precision, recall and accuracy. Variance over 10 iterations is shown in brackets.

Ranking	Intention Labels
0.33	A-CW
0.21	A-GAP
0.08	A-DESC
0.07	BG-NE
0.05	A-USE
0.04	BG-EP
0.00	CW(-+)
-0.01	BG(-+)
-0.01	TXT
-0.04	CW-DESC

Table 8.3: Author intention labels ranked in terms of importance-Logistic Regression

marginally less variation in runs, although there is no significant difference between SVM and LLR. Accuracy between SVM and LLR is significantly different from that of the decision tree method. One of the reasons for the latter's poor performance may be that the label features are not exclusive. For example, although author gap and author/cited differences (A-GAP, A-CW) are rare in Poor examples, they are not completely absent.

There are no direct systems to compare to but Criterion (cf. Section 2.3.1.3) algorithm details described in (Burstein et al., 2004) show agreement between the system and human score of essays at 97%. This is, however, a commercial system built on multiple elements, not just author intentions and a much larger training set. Whilst this level of accuracy is not achieved the results are promising as a first step, and with the addition of other features accuracy could be improved. For example, experimenting with adding sentence counts and citation counts consistently improved the accuracy by 4%.

Table 8.3 ranks labels in terms of importance using the Logistic Regression classifier, showing that an author highlighting a difference of their work to a cited work or how their work addresses a gap are the most important labels for distinguishing between quality ratings. This seems plausible given earlier discussion in Section 2.5 about problems with *Related Works*, e.g. Maxwell (2006) who states that cited work needs to be shown to have implications for the study, and the findings in Section 3.8 about what experts expect to see in a *Related Work*. It seems that if this type of connection is missing, then the work is rated as poorer.

Finally, for the best performing model SVM, the confusion matrix is considered. Here, the interest was to see if mis-classification was occurring in the nearest group, i.e. Good were mis-classified as Fair and not Poor. Out of 10 iterations, this happened twice – one Poor *Related Work* was classified as Good – and 6 times one Good *Related Work* was classified as Poor. We looked at each of these to understand why this might be. The Poor *Related Work* that was mis-classified as Good was short with several CW(-+) evaluation sentences but contained only one sentence, A-USE, that referred to the author's work. This was likely judged by the annotator as poor because the relevance to the author's work was not clear; they just said they used another cited work. The Good *Related Work* that was misjudged as Poor was a very long *Related Work* with a large number of label intentions of CW-DESC. There were, however, several comparisons of the au-

thor's work to the cited work. The mis-classification is related to the occurrence of labels.

8.8 Discussion and Conclusions

Using author intentions developed in this thesis, the results show that some author intentions differ significantly across *Related Work* sections rated Poor, Fair and Good. These findings confirm earlier discussion that poorer *Related Work* sections will contain bland cited work descriptions with no context to the author's work or identification of the gap that is being filled. The prediction of quality rating is consistently accurate at 70% with only author intentions as features. While this does not match commercial tool accuracy, such as Criterion (97%), it is a very promising result despite the limitation of small sample size. Overall, the author intentions show promise as being viable indicators of quality of the content and demonstrate that the author intentions model developed is a good representation of the expected content in a *Related Work*.

As seen in other systems that predict scores, the classifier could be improved by including more features. In particular, one area that could prove insightful to improving the performance of the classifier could be in studying patterns of labels occurring together. When observing the mean occurrence and variance of labels in Table 8.1, it was not simply a case of a Poor section not having any sentences about the author's work or never mentioning a gap. There may be more to learn about patterns that happen with labels occurring together that support the better classification of the different ratings.

Reaching human level of judgement for peer-review in scientific papers is most likely impossible. For example, it is hard to tell what is missing, specifically what has not been addressed or identify something that is incorrect – these aspects might still require a human expert. We saw evidence of this in Section 3.7.2 when our PG students could not tell that cited works were not relevant given the *Introduction* compared to experts. Nonetheless, this type of quality rating, if developed at a section specific level, could prove useful in tasks other than writing support, such as supporting peer-review, directing where reviewers time should be focused and on which papers. The highlighting of salient content sentences has been shown to assist peer-review to filter bad papers more quickly (Sándor

and Vorndran, 2014). In addition, it could help novice readers better interpret the content of what they are reading; an aspect highlighted by some of the users in evaluation **LitCrit** in Section 6.5.2. These aspects, however, would all need further evaluation.

Chapter 9

Discussion and Conclusions

9.1 Summary of Contributions and Results

This thesis makes two main contributions: (1) It uses peer-review to understand what content should be present in *Related Work* and builds a model of author intentions to represent this content, using it to support writing feedback; and (2) it demonstrates that this author intention model can be reliably annotated by humans and builds a classifier to reliably automate the recognition of these intentions within a *Related Work*.

These contributions fall into four areas. We summarise our results based on these four areas, our research questions and discuss the degree to which our hypotheses are supported.

9.1.1 Building an Author Intention Model to Support *Related Work*

We hypothesised that if a reasonable agreement could be found between experts during peer-review, then peer-review could be used to understand what content experts expect to see in a *Related Work*. Additionally, peer-review could be used to compare PG students to experts to understand better where PG students struggle. We also hypothesised that the model of intentions built from expectations of content should represent a quality measure of a *Related Work*, and this could be tested by using the intentions to predict quality.

The research questions in relation to this were:

1. What are the content expectations highlighted in a *Related Work* by experts and do experts agree with each other?
2. Do PG students differ from experts in what they look for in a *Related Work*?
3. Do the author intention labels serve as a proxy for indicating content quality in a *Related Work*?

Our contribution is to show that peer-review can be used to understand what content should be present in *Related Work*. The problem of agreement on content during peer-review is widely known, and our approach to using peer-review is in contrast to other work which use observational studies of published papers. Our work reveals that there is an agreement in what experts look for when assessing a *Related Work* during peer-review. In addition, by comparing expert and PG student peer-review, we find particular areas that PG students struggle with. This is demonstrated by the differences between these two groups in their qualitative and quantitative responses in our user study about aspects that are present or missing in *Related Works*. From these findings of what experts look for in content and which areas PG students struggle, we build an author intention model to represent content expected within *Related Work*. We validate our model of author intentions as a good proxy of expectations of content when we show it can be used to predict the quality of a *Related Work* with good accuracy. When intentions are missing from *Related Works*, those *Related Works* can be shown to be of lower quality.

Our results support our hypotheses, but there are limitations to this (discussed in Section 3.8.1): our sample size and numbers of *Related Works* peer-reviewed was small, and possible bias existed in our convenience sampling method. Therefore, the results need to be considered under these limitations, and any future work may want to consider the suggested improvements in Section 3.8.1. However, as pointed out in Section 3.8.1, the findings are not that dissimilar to what might be expected from observational studies of *Related Work* or observations on PG student work in terms of where they struggle, suggesting the approach is viable, and our hypotheses supported.

9.1.2 An Effective Related Work Feedback Tool

We hypothesised that the highlighting of narrative could influence thinking and could be measured by comparing PG student responses during peer-review with and without this highlighting of the narrative. The research questions in relation to this were:

1. Does highlighting the narrative structure with intentions change PG student perceptions of a *Related Work*?
2. Do PG students find the visualisation of intentions and feedback on missing aspects helpful?

Our contribution is **LitCrit**, a writing analytic tool developed as part of this thesis, which highlights the narrative of a *Related Work* with author intentions. We carry out a further study with PG students, and find evidence that using **LitCrit** influences PG student thinking about the *Related Work* as they now identify aspects they previously missed during peer-review. Using **LitCrit** results in significant changes in context and quality ratings, which brings PG students' ratings in line with experts. Our evaluation shows that overall, the PG students find the highlighting of intentions and provision of the feedback in **LitCrit** useful.

Whilst our results show support for our hypotheses, we must consider the limitations of the study – described in Section 6.6.1 – about sample size, learning effect and potential bias of our subjects. We describe a number of steps that could be taken in future work, such as reducing the time lag between the two studies to ensure no learning effect is present, or different forms of evaluation of **LitCrit** that could provide further validation of findings.

9.1.3 Approach to Automating Recognition of Author Intention in a *Related Work*

We hypothesised that an approach based on feature engineering would produce results that could support author intention recognition. However, more importantly this approach allowed us to explore errors and understand aspects important for feedback or that relate to a better understanding of pedagogical implications.

The research questions were:

1. Can the author intention labels be annotated with reasonable human agreement?
2. Can the author intention labels be recognised automatically with reasonable accuracy?

Overall, our hypothesis is supported by our results, as we show that humans can reliably annotate our author intention model. Additionally, we build a model based on supervised machine learning to recognise author intentions, which achieves accuracy within 0.60% of human annotation performance. Our error analysis, through looking at incorrectly classified labels and features, gave insight to the cause of errors, and raised some pedagogical issues.

In addition, in order to support providing feedback within **LitCrit**, we propose and evaluate a method to segment and label the discourse of *Related Work*. Our method creates contiguous blocks of sentences (segments) that are contextually similar, i.e. discussing local topics, such as a citation, how the author's work differs, background work. We use the author intention labels to decide on labels for segments. The labelling of segments is what we use to understand the context of the *Related Work* and provide overall feedback in addition to highlighting author intentions within **LitCrit**. We show our method has very high accuracy at segmenting and at choosing correct descriptive labels for segments.

9.1.4 Discipline Independence of Model

Finally, we investigate the discipline independence of the author intention model, applying it to the discipline of Computer Graphics. The performance in this discipline is lower than in the Computational Linguistics discipline. However, we show how focusing on features that are less domain-specific, e.g. the removal of our label suggestion feature, and considering the different structure intentions between disciplines within the *Related Work* by removing features related to structure, e.g. the previous label feature, can improve performance.

We do note in Section 7.2 that our approach does not take advantage of current state-of-the-art transfer learning such as pre-trained models. This is because we took the decision to take a feature engineered approach, but we discuss this more in our suggestions on how to improve the NLP in Section 9.2.3.

9.2 Insights, Limitations and Future Work

9.2.1 LitCrit, Limitation and Future Improvement

We have shown that our model of author intentions represents the content in a *Related Work*, and the highlighting of these intentions does influence and draw attention to aspects of writing PG students previously missed. However, this work does not address the question of whether using **LitCrit** can improve PG students' writing. Answering this question requires more work in developing **LitCrit**.

In order to fully deploy **LitCrit** automating the identification of co-referencing needs to be addressed. Error analysis carried out on both the predictions made by our supervised classifier and in segmenting the discourse for feedback showed that accurate identification of co-referencing would be a critical component of **LitCrit**. While there has been work in this area with reasonable success in academic writing, (Rösiger and Teufel, 2014; Batista-Navarro and Ananiadou, 2011; Gasperin, 2009) it will still introduce an element of error.

We found during peer-review that experts and PG students both look for the content to be clear with good structure and presentation. However, our work did not consider any of these areas, but they are all essential aspects of writing and may impact **LitCrit's** performance when it is fully automated. It may be necessary to implement features currently found in tools such as Grammarly[1] or Turnitin[2] to mitigate against these issues.

Human agreement of intentions in academic writing is known to have problems due to subjectivity and the knowledge brought by the subject. The annotation study carried out showed good human agreement, but still, there was some disagreement on labels. We observed similar differences between our annotators and our classification errors with labels such as A-GAP/A-DESC and CW-DESC/BG-EP. Often either label could be argued as correct. PG Students highlighted issues with sentence label accuracy when evaluating **LitCrit**, but they also said disagreeing with a label was not necessarily a bad thing as this made them challenge their thinking. This aspect requires further evaluation in

[1] http://www.grammarly.com
[2] https://www.turnitin.com

the future to understand how the accuracy impacts the feedback interpretation by the student, or if it makes them think deeper about their writing.

Currently, **LitCrit** feedback only identifies what segment labels are present with two possible further comments on whether context to the author's work is present or if there is critical evaluation of cited work. Examples of feedback presented in Section 6.3.2.2 demonstrates that the feedback can be quite dry and repetitive. The first feedback example gave some suggestions about the writing based on the low count of intention labels that indicated evaluation, and the low count of intention sentences that indicated discussion of the author's own work. However, the second example of feedback just reiterated what was there as minimum counts had been met. While some PG students in the evaluation thought the feedback was useful, others thought it was more useful for novices, and just reiterated what the labels already highlighted. A fully automated version needs to provide valuable feedback, not just a reiteration of what has been highlighted by the sentence labelling. Thus, more work is needed on how the information about the intention labels and context of segments can be combined to generate feedback. Our approach to generating feedback is similar to work done in the field of natural language generation (NLG) which produce text from numerical data using templates, e.g. (Isard and Knox, 2016). We observed that the mean occurrence of labels differs from poor and good *Related Work* and further study could learn more informative feedback, e.g. relating low occurrence of label types to the likelihood more discussion is needed in a particular area. Work in NLG has also used uncertainty in data, for example, probabilities about the likelihood of weather, showing that these can be put in textual summaries to aid human decision making (Gkatzia et al., 2016). Having more information about label occurrence and its relation to quality may help to explore how these types of NLG methods could be exploited to improve feedback.

9.2.2 Other Uses for LitCrit

This approach of focusing on fine-grained intention in one section could be applied to other sections of a research article to develop content models for these sections. However, there is potential for other uses. The highlighting of intentions when using **LitCrit** was an aspect PG students commented could help when reading articles. Combining predictions of quality and highlighting of au-

thor intention may have future use in helping readers understand their reading material or in helping during peer-review. Drawing attention to aspects during peer-review can have a positive influence. For example, Sándor and Vorndran (2014) show that content-orientated highlighting of sentences in an article during peer-review helps focus a reviewer's attention, promoting the rapid filtering out of bad papers.

9.2.3 Improving Aspects of NLP

Like other models that classify intentions, we find infrequent labels are harder to predict and improvements for individual label prediction is needed. Particularly problematic are (i) descriptions between sentences that talk about the author's work in general and those that expose the novelty of the work; (ii) sentences that provide a critical evaluation of a citation or the background/field.

We highlighted the first problem in the previous section, pointing out it is important to understand what impact these type of errors have on the PG students interpretation when using the system. This understanding will help focus on where or if improvement is needed here. Looking at how and why labels were mis-classified some insight was gained about the second problem - classifying critical evaluation labels. Evidence often showed in these sentences, words or phrases were not frequent enough to be in the lexicon. This means that the identification of the critical evaluation within the sentence did not occur.

Recent work offers some possibilities on how features could be enhanced to improve labelling. Firstly, patterns we used to augment our features, described in Section 5.9.2, were manually derived. Other work has shown that using bootstrapping techniques to augment a pattern set can prove successful, e.g. Jurgens et al. (2018) generated over four times that of manually curated patterns. They identified new patterns that apply not just based on one sentence but also look at the preceding or following sentences. This type of approach might help expand the limitation of the current pattern set. Secondly, the use of WordNet roots for Nouns, e.g. where nouns are taken to their more general form (e.g., *mm* and *cm* become *quantity*), has been shown as a useful feature for author intention identification (Asadi et al., 2019). This application of WordNet is one possible avenue that may assist in both transitioning the pattern list to another domain but also

expanding the lexicon to capture variation found in critical evaluation labels.

Overall a solution that would help address both issues would be to have access to a more substantial amount of annotated data, but creating annotated data is time-consuming and expensive. We specifically chose a hand-engineered approach to explore features and errors generated and gain a level of understanding of how the features influence the classification. We saw this as beneficial in understanding any possible pedagogical implication to enhance writing support. Recent advances in NLP with pre-trained models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), discussed in Sections 2.7 and 5.4, highlight that our results using a feature approach were likely sacrificing interpretability for state-of-the-art performance. In addition, as discussed in Section 7.2 not using these methods limits our ability to make our model generalisable and domain independent. Future work in this area should now consider the use of pre-trained word embeddings as features to the model and what pre-trained models, such as those using SciBERT (Beltagy et al., 2019) could bring.

9.2.4 Pedagogical Insight to Support PG Writing

In Section 2.5, we discussed the kind of difficulties PG students have when writing about *Related Work* and we found evidence to support the ideas discussed. Our findings in Section 3.7.2 overall pointed to PG students being less likely to notice that context between the author's work and cited work was missing. While using **LitCrit** did draw a PG student's attention to citation evaluation, it still resulted in differences in ratings between the students and experts, although not significant. Students seem to have a different understanding of citation evaluation purpose to experts with students focusing more on limits and merits rather than how the context of this evaluation was needed. They also differed to experts in rating the importance of critical evaluation, seeing it as less important than experts. This indicates a more profound misunderstanding of the function of critical evaluation and may require pedagogical intervention or more instructional information to support writing. We also saw evidence that students were likely to be misled by superficial aspects, such as the engaging writing in *Related Work* F (cf. Section 3.7.2) and not realise those essential aspects were missing until they reviewed *Related Work* F using **LitCrit**.

In Section 2.6 we drew attention to a study by (Jurgens et al., 2018), which we feel offers insight into another aspect that may contribute to why PG students struggle writing *Related Work*. This is the idea that the decline in recent NLP papers of discussion that offers positioning or comparison to other work is influencing novices. Perhaps this results in these behaviours of writing perpetuating as novices see this as the norm in the materials they read. We also believe the significant growth in the NLP field and focus of research in recent years raises another valid question - How much of this change is due to pressures to publish and people having less time to spend on quality writing? Less writing time could result in the aspects that are more difficult to write, or that writers assume their audience will know of being omitted. While these aspects are somewhat speculative, nonetheless, they may be worthy of more investigation as to how much influence they are contributing to writing standards.

Appendix A

Appendix A - List of Related Work Papers Used in Study

Paper A

Name : Identification and Resolution of Chinese Zero Pronouns: A Machine Learning Approach

title=Identification and resolution of Chinese zero pronouns: A machine learning approach,

author=Zhao, Shanheng and Ng, Hwee Tou,

booktitle=Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),

pages=541–550,

year=2007

Paper B

Name : Multilingual Harvesting of Cross-Cultural Stereotypes

title=Multilingual harvesting of cross-cultural stereotypes,

author=Veale, Tony and Hao, Yanfen and Li, Guofu,

booktitle=Proceedings of ACL-08: HLT,

pages=523–531,

year=2008

Paper C

Name: Extraction of Entailed Semantic Relations Through Syntax-based Comma Resolution

title=Extraction of entailed semantic relations through syntax-based comma resolution,
author=Srikumar, Vivek and Reichart, Roi and Sammons, Mark and Rappoport, Ari and Roth, Dan,
booktitle=Proceedings of ACL-08: HLT,
pages=1030–1038,
year=2008

Paper D

Name : Generating Lexical Analogies Using Dependency Relations

title=Generating lexical analogies using dependency relations,
author=Chiu, Andy and Poupart, Pascal and DiMarco, Chrysanne,
booktitle=Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),
pages=561–570,
year=2007

Paper E

Name : Automatic Image Annotation Using Auxiliary Text Information (P08-1032)

title=Automatic image annotation using auxiliary text information,
author=Feng, Yansong and Lapata, Mirella,
booktitle=Proceedings of ACL-08: HLT,
pages=272–280,
year=2008

Paper F

Name : Using Foreign Detection to Improve Parsing Performance
title=Using foreign inclusion detection to improve parsing performance,
author=Alex, Beatrice and Dubey, Amit and Keller, Frank,
booktitle=Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),
pages=151–160,
year=2007

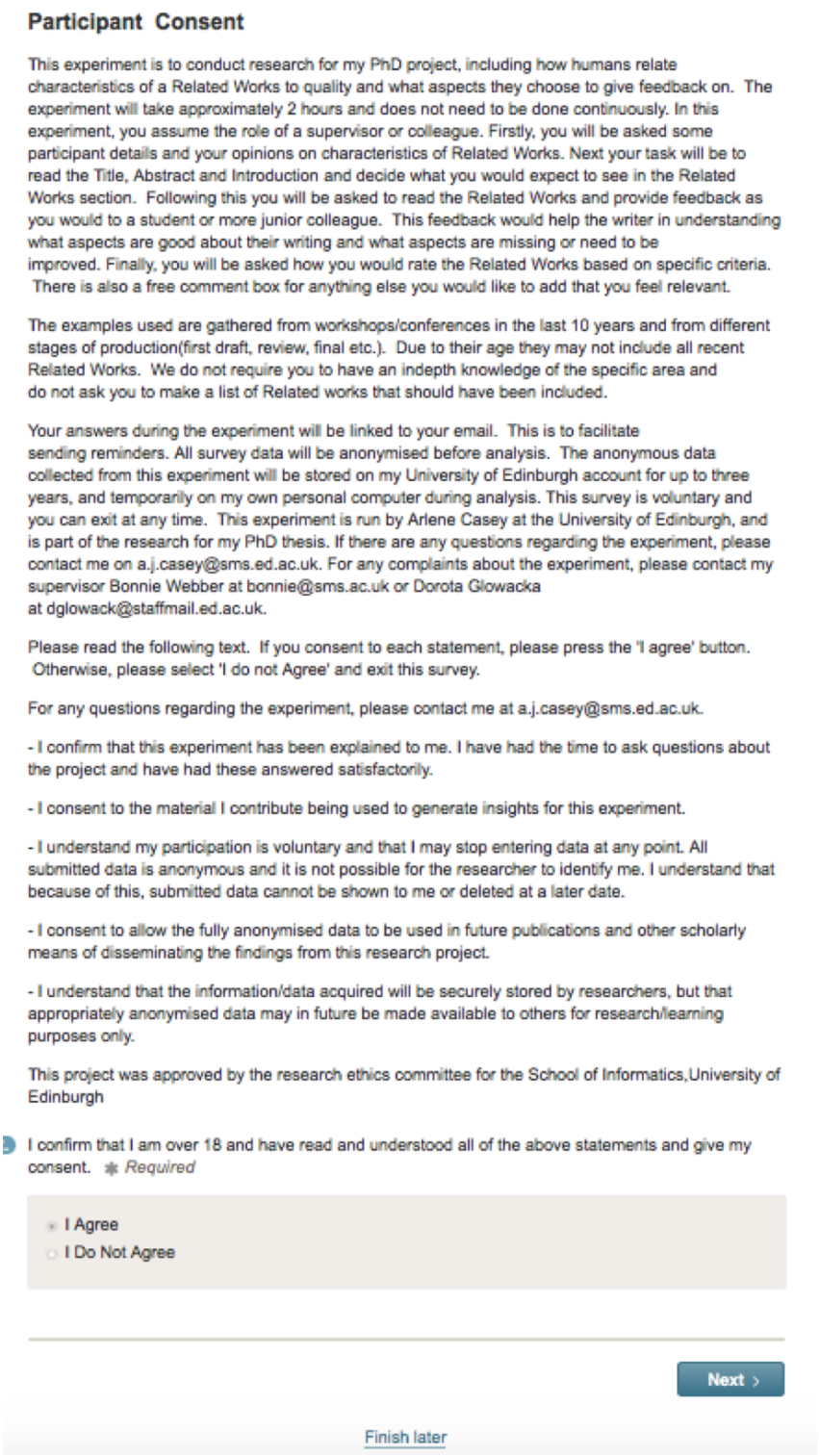
Paper G

Name : A Re-examination of Query Expansion Using Lexical Resources
title=A re-examination of query expansion using lexical resources,
author=Fang, Hui,
booktitle=proceedings of ACL-08: HLT,
pages=139–147,
year=2008

Appendix B

Appendix B - Screenshots of Questions Asked During User Studies

This appendix shows screenshots of questions asked during user study 1, described in Chapter 3 and user study 2, described in Chapter 6.



Participant Consent

This experiment is to conduct research for my PhD project, including how humans relate characteristics of a Related Works to quality and what aspects they choose to give feedback on. The experiment will take approximately 2 hours and does not need to be done continuously. In this experiment, you assume the role of a supervisor or colleague. Firstly, you will be asked some participant details and your opinions on characteristics of Related Works. Next your task will be to read the Title, Abstract and Introduction and decide what you would expect to see in the Related Works section. Following this you will be asked to read the Related Works and provide feedback as you would to a student or more junior colleague. This feedback would help the writer in understanding what aspects are good about their writing and what aspects are missing or need to be improved. Finally, you will be asked how you would rate the Related Works based on specific criteria. There is also a free comment box for anything else you would like to add that you feel relevant.

The examples used are gathered from workshops/conferences in the last 10 years and from different stages of production(first draft, review, final etc.). Due to their age they may not include all recent Related Works. We do not require you to have an indepth knowledge of the specific area and do not ask you to make a list of Related works that should have been included.

Your answers during the experiment will be linked to your email. This is to facilitate sending reminders. All survey data will be anonymised before analysis. The anonymous data collected from this experiment will be stored on my University of Edinburgh account for up to three years, and temporarily on my own personal computer during analysis. This survey is voluntary and you can exit at any time. This experiment is run by Arlene Casey at the University of Edinburgh, and is part of the research for my PhD thesis. If there are any questions regarding the experiment, please contact me on a.j.casey@sms.ed.ac.uk. For any complaints about the experiment, please contact my supervisor Bonnie Webber at bonnie@sms.ac.uk or Dorota Glowacka at dglowack@staffmail.ed.ac.uk.

Please read the following text. If you consent to each statement, please press the 'I agree' button. Otherwise, please select 'I do not Agree' and exit this survey.

For any questions regarding the experiment, please contact me at a.j.casey@sms.ed.ac.uk.

- I confirm that this experiment has been explained to me. I have had the time to ask questions about the project and have had these answered satisfactorily.

- I consent to the material I contribute being used to generate insights for this experiment.

- I understand my participation is voluntary and that I may stop entering data at any point. All submitted data is anonymous and it is not possible for the researcher to identify me. I understand that because of this, submitted data cannot be shown to me or deleted at a later date.

- I consent to allow the fully anonymised data to be used in future publications and other scholarly means of disseminating the findings from this research project.

- I understand that the information/data acquired will be securely stored by researchers, but that appropriately anonymised data may in future be made available to others for research/learning purposes only.

This project was approved by the research ethics committee for the School of Informatics,University of Edinburgh

I confirm that I am over 18 and have read and understood all of the above statements and give my consent. * *Required*

- I Agree
- I Do Not Agree

Next >

Finish later

Figure B.1: The figure shows the initial consent and instructions screen given to participants in user study 1, Chapter 3.

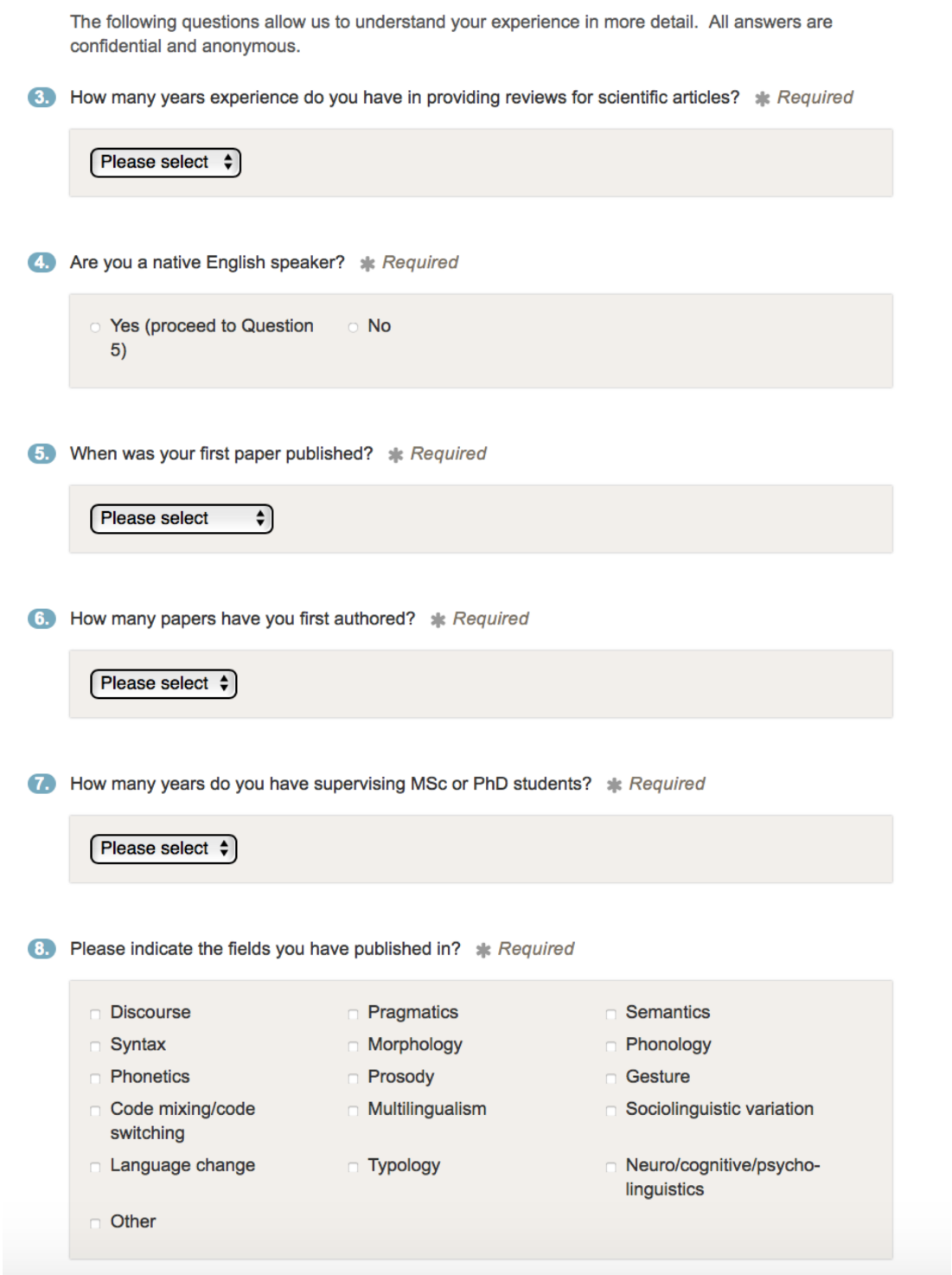


Figure B.2: The figure shows the demography questions asked of participants in user study 1, Chapter 3.

Figure B.3: The figure is a continuation of the demography questions asked of participant in user study 1, Chapter 3, continued from the previous page.

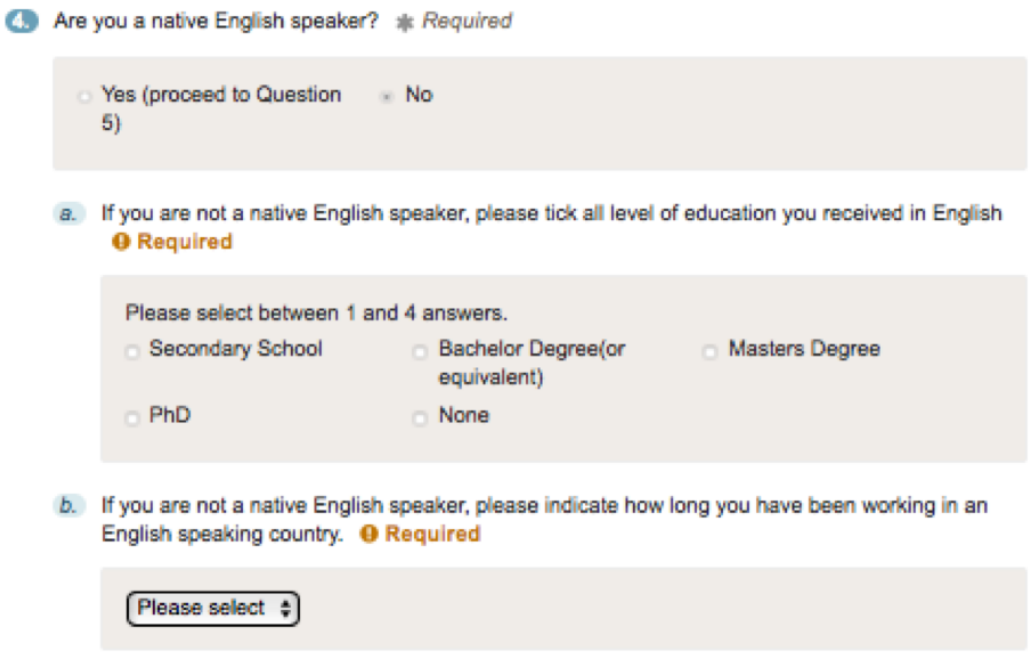


Figure B.4: The figure shows the additional questions for participants who were not native English speakers in the demography question section of user study 1, Chapter 3.

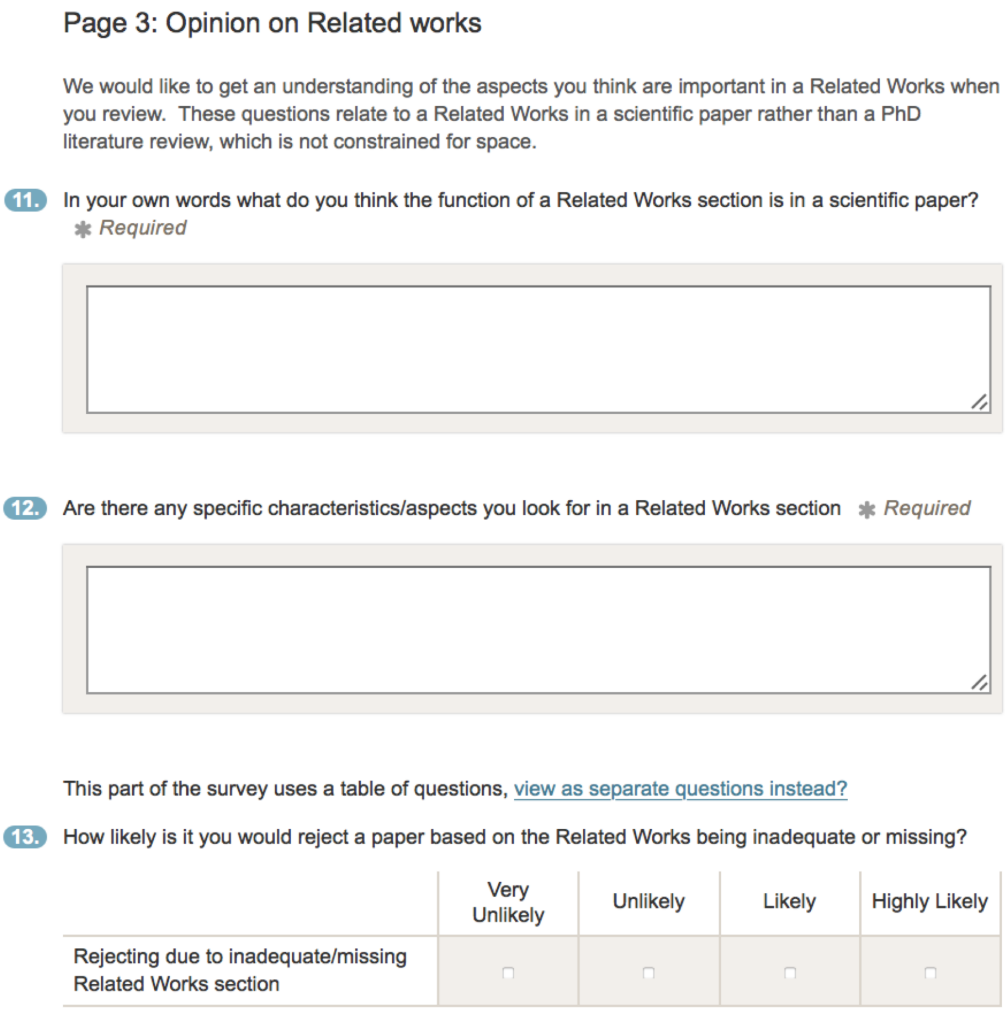
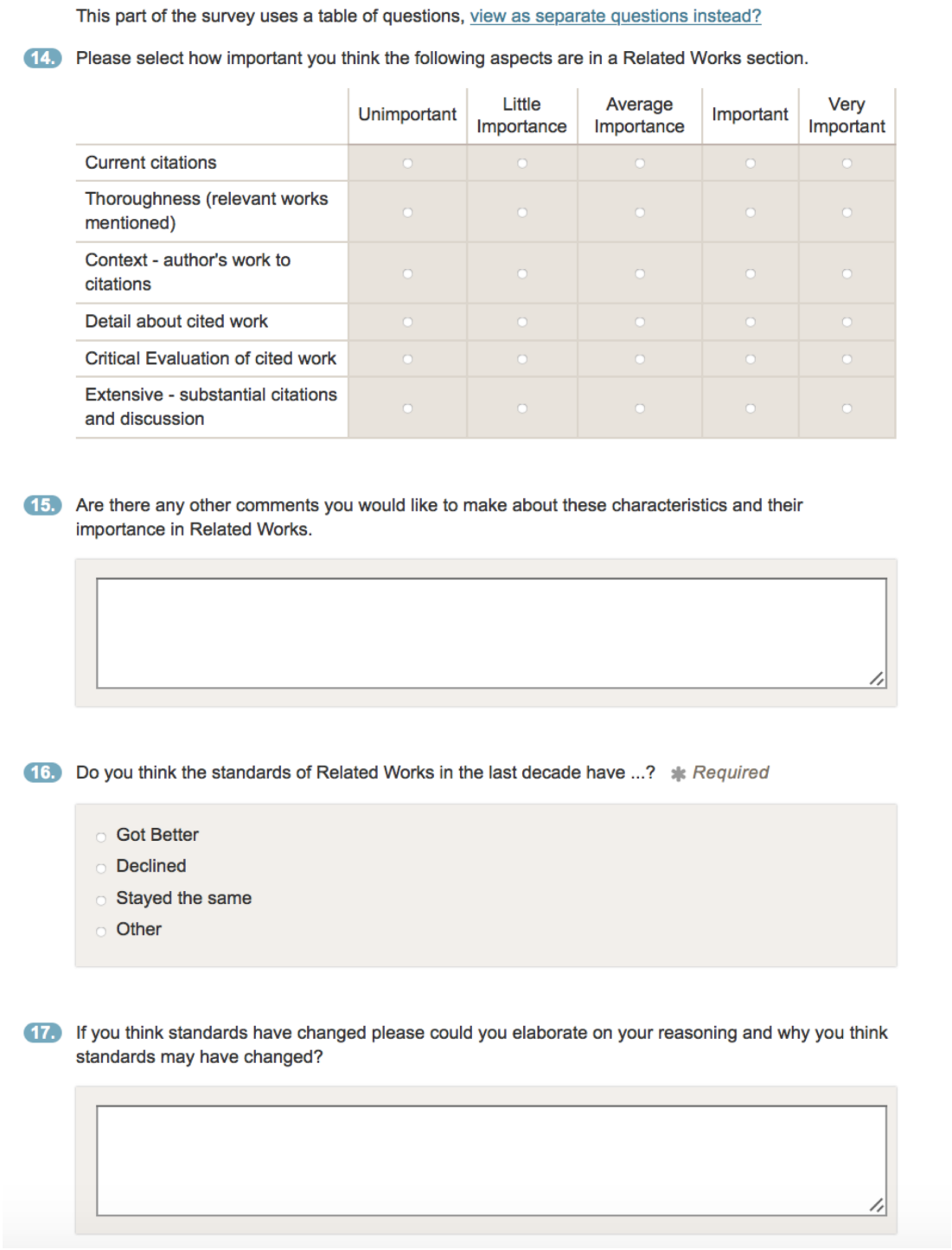


Figure B.5: This figure shows the questions asked of participants about function and characteristics participants look for in a *Related Work* in users study 1, Chapter3



This part of the survey uses a table of questions, view as separate questions instead?

14. Please select how important you think the following aspects are in a Related Works section.

	Unimportant	Little Importance	Average Importance	Important	Very Important
Current citations	○	○	○	○	○
Thoroughness (relevant works mentioned)	○	○	○	○	○
Context - author's work to citations	○	○	○	○	○
Detail about cited work	○	○	○	○	○
Critical Evaluation of cited work	○	○	○	○	○
Extensive - substantial citations and discussion	○	○	○	○	○

15. Are there any other comments you would like to make about these characteristics and their importance in Related Works.

16. Do you think the standards of Related Works in the last decade have ...? * *Required*

- ○ Got Better
- ○ Declined
- ○ Stayed the same
- ○ Other

17. If you think standards have changed please could you elaborate on your reasoning and why you think standards may have changed?

Figure B.6: The figure show the questions asked of participants about aspects of *Related Works* in user study 1 Chapter 3.

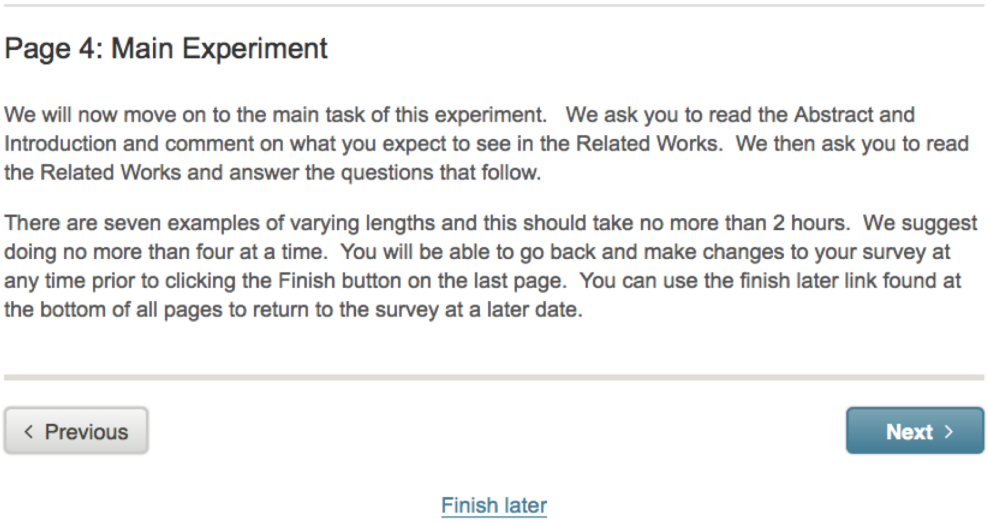


Figure B.7: The figure show the instructions given to participants as they move to the main task of peer-review in user study 1 Chapter 3.

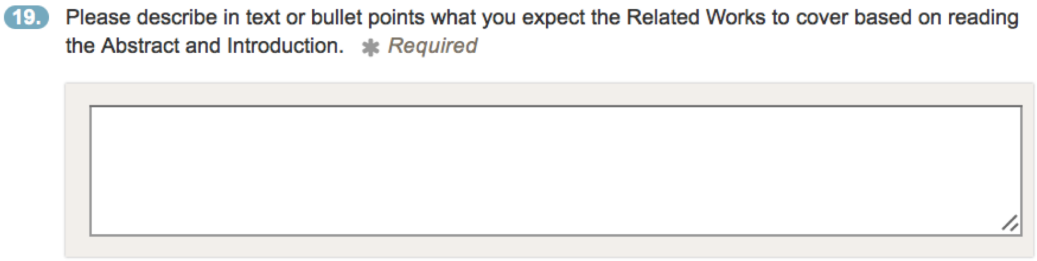


Figure B.8: The figure shows the question which asks participants to think about aspects they expected the *Related Work* to contain in user study 1 Chapter 3.

LitCrit- Experiment 2

Instructions

Like last time you will be asked to read the Abstract, Introduction and then the Related Work section for each paper - there are 7. This time the Related Work section will be presented via the **LitCrit** website. On loading the Related Work it will be provided in normal text without any feedback mark-up. Please use the Feedback tab - described below- to put the Related Work in 'feedback' mode. We provide both modes as we realise some people may prefer to read the text then look at the mark-up text. In order for our experiment to work we do need you to view the 'feedback' mode before scoring the Related Works. In this experiment we do not ask you for any free-text only multiple-choice ratings. We do provide a free-text box after each set of multiple choice ratings - it is optional but please use it if you feel you want to comment on anything at all.

LitCrit focuses on content we expect to see and not on style of writing punctuation or grammar. However, the feedback is auto-generated and may guess incorrectly so please feel free to disagree. **LitCrit** is not a knowledge-base and does not know about existing literature that you as a human may know is missing nor can it tell if the Related Work is off topic.
There are three tabs on the **LitCrit** web page
(i) '**Original Text**' which shows the original Related Work with no markup,
(ii) '**Feedback**' which shows the Related Work marked up as described below
(iii) '**About**' which brings you back to this information.

Each sentence of the Related Work is color coded into one of three categories :
Background
Cited Work
Author

Background (BG) sentences provide general information about the field or common knowledge, they may have citations but usually these are provided as evidence and not part of the sentence syntax. Cited Work (CW) sentences are about a specific cited work. Author (A) sentences are about the author's work in this paper you are reading. In addition, each sentence is labelled as to its specific purpose within the category of Background, Cited Work or Author. This label is found at the start of the sentence e.g (BG-E BG-NE) You can hover over these and a pop-up describes the label.
To the right of the Related Work text you will find a list of the labels - you can hover over these to for a pop up explanation. To the right of the text there is also a comments box about the Related Work.

Figure B.9: The figure show the instructions given to participants in the second user study in Chapter 6.

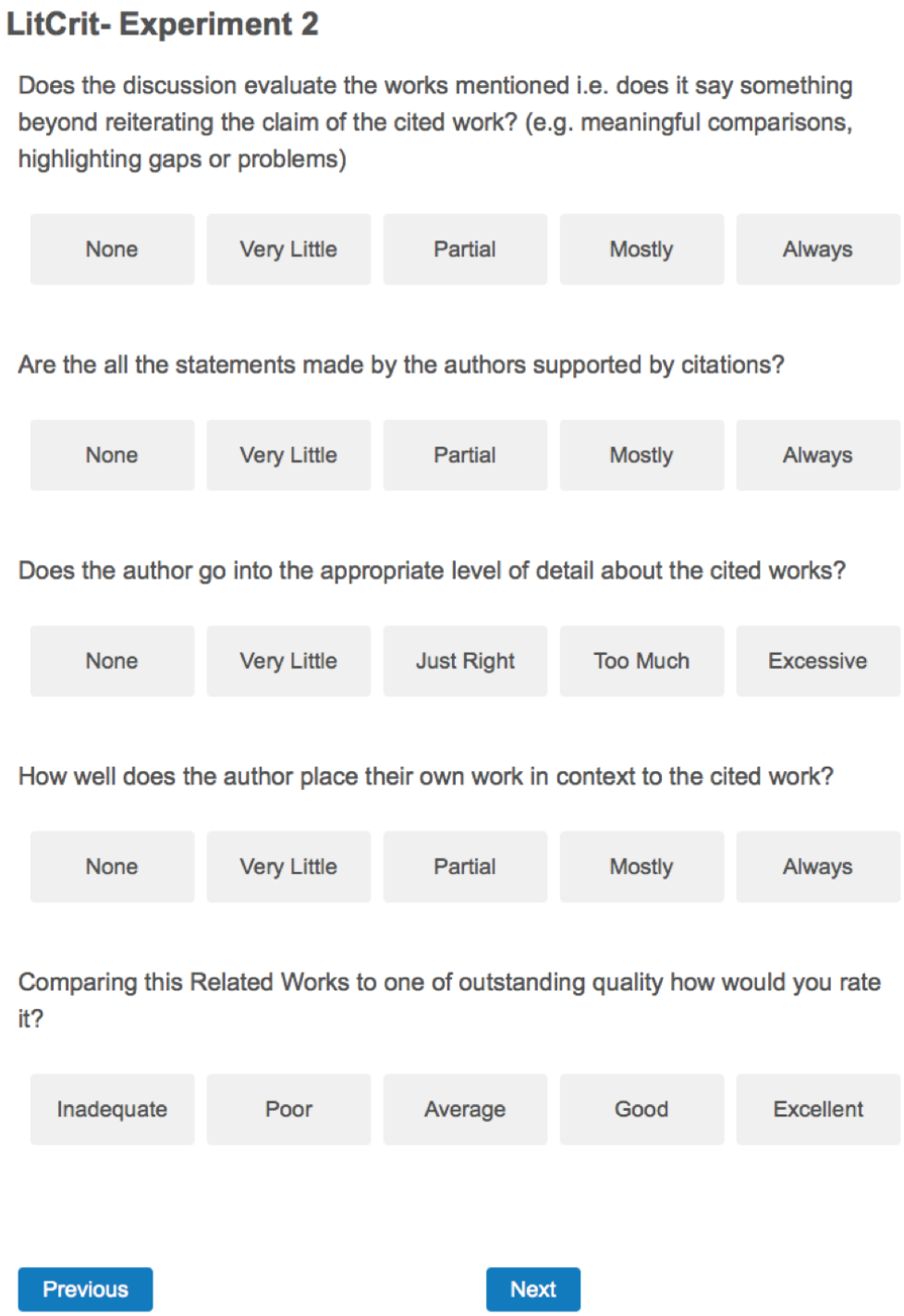


Figure B.10: The figure show the questions asked in user study two Chapter 6 after a *Related Work* was read by the participant.

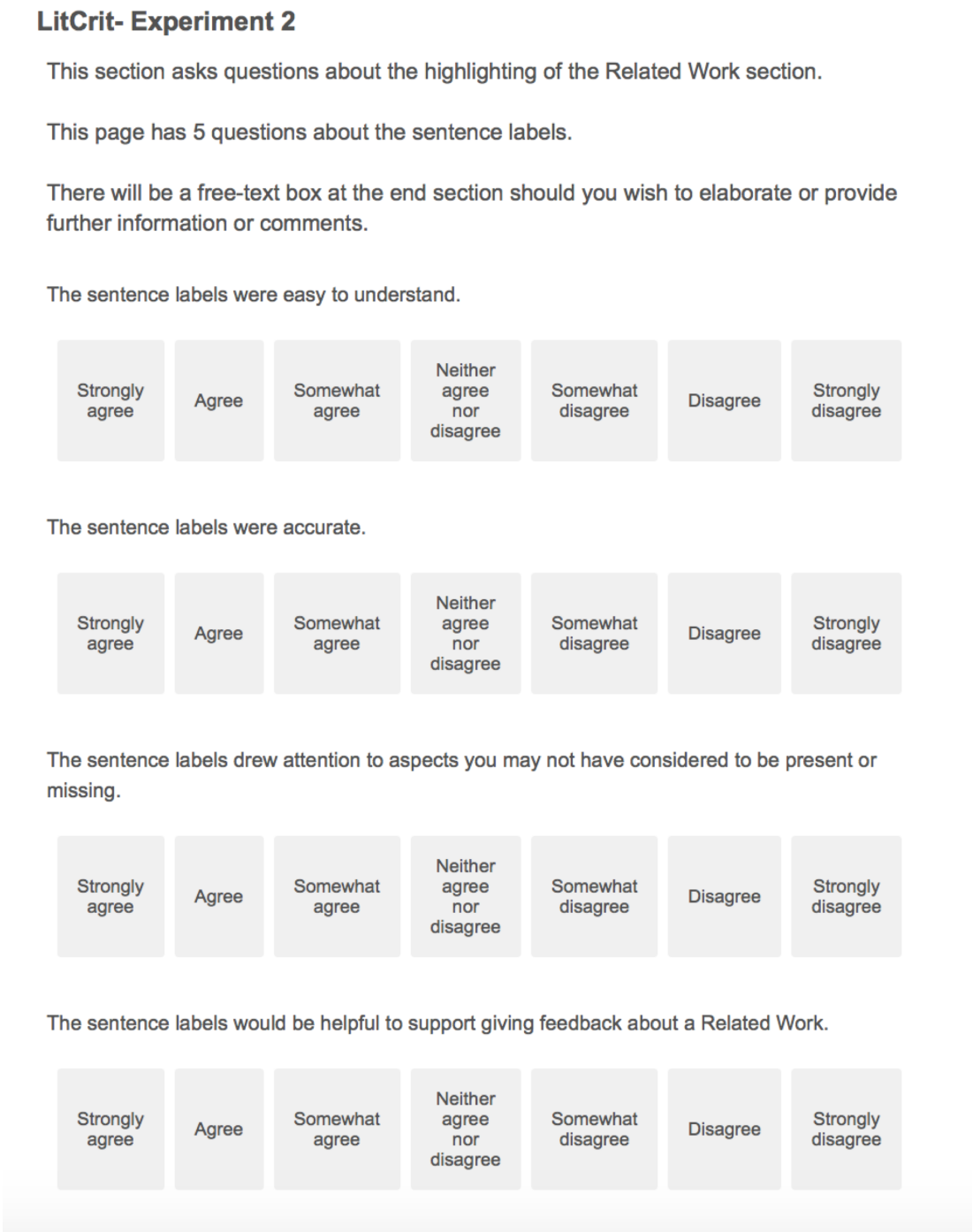


Figure B.11: The figure show the questions asked in user study two Chapter 6 to evaluate **LitCrit** sentence intention labelling following reviewing of all *Related Works*.

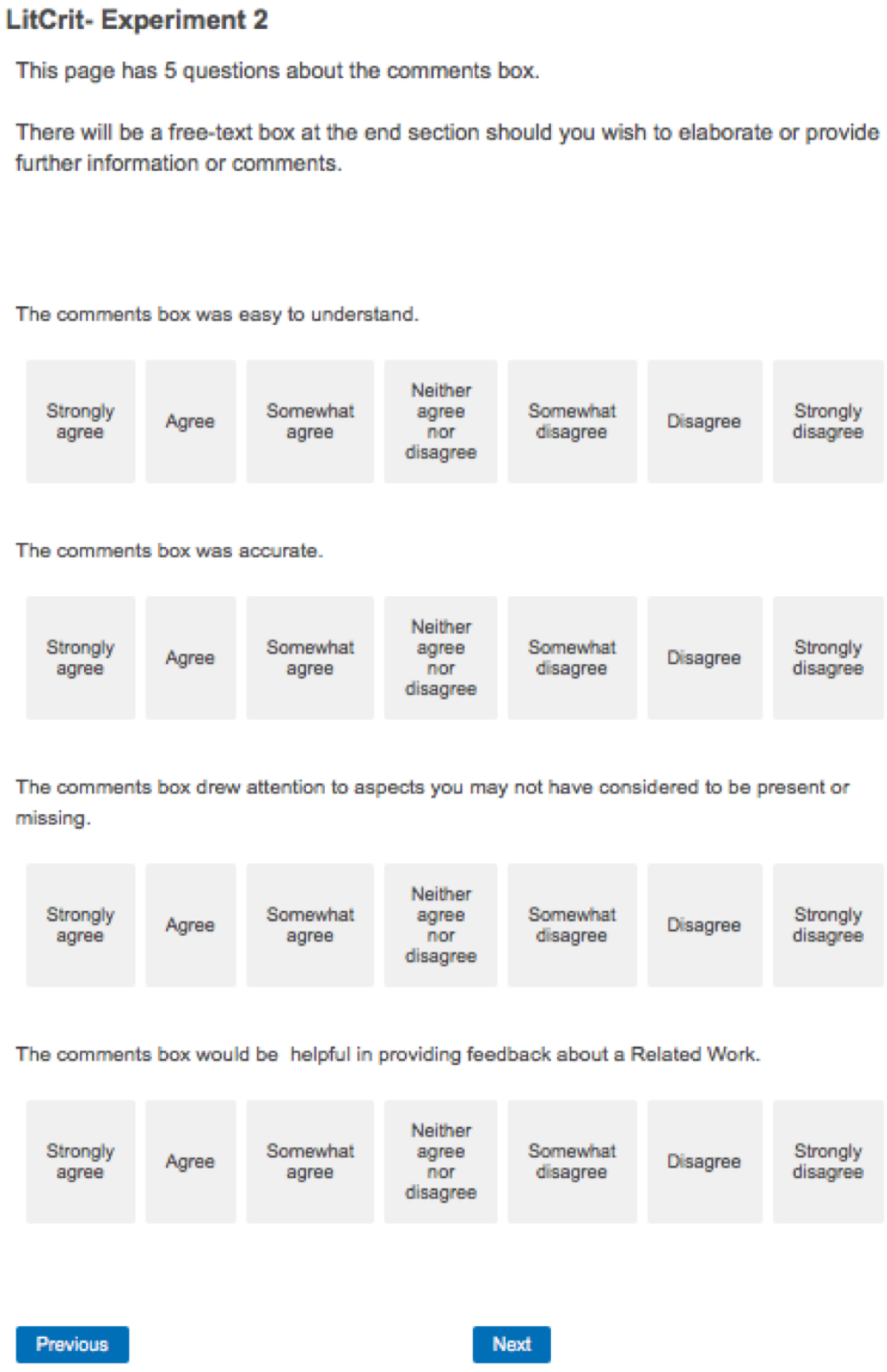


Figure B.12: The figure show the questions asked in user study two Chapter 6 to evaluate **LitCrit** feedback comments following reviewing of all *Related Works*.

Appendix C

Appendix C - Annotation Guidelines

Annotation Guidelines – Labelling Related Works Sentences

The purpose of this document is to give instructions for annotating sentences for Related Works.

All of the papers have been published in Conference Proceedings in Computational Linguistics. They were downloaded from the ACL Anthology (http://www.aclweb/anthology). All papers are between 7 and 9 pages long. The related work sections have been captured by pdf to text and OCR recognition and there may be some missing words or letters.

Please read the guidance fully before carrying out any annotation.

Annotation Guidance

The procedure involves reading each sentence and assigning a label to the sentence. This will be done within Excel. The spreadsheet has five columns:

- Document ID (pre-populated)
- Sentence ID (pre-populated)
- Original Sentence (pre-populated)
- Marked up Sentence (pre-populated, this has placeholders for citations and coreferencing). This is included as sometimes it is easier to read if there are many citations.
- Annotation Label (dropdown selection only)
- Comments – free-text to allow you to write notes e.g. there is a problems with this sentence and you were unsure

The **Label** column has been pre-populated with a drop down of all annotation labels. In the next section we will discuss the annotation labels and give examples.

Determining Labels

In order to determine a label for the sentence first determine the subject of the sentence:

Is the sentence about:

- Field in general, general knowledge, a group of cited works(previous works)
- Other's work (cited work)
- Author's work
- Multi-type subject - does it discuss/compare any of the above
- Text – structure information about what is coming next

GUIDELINE FOR DECIDING THE SUBJECT

Field in general, background knowledge

Here the author may be making general assertions that are known about the field, providing citations as evidence or not or they could describe work in general terms. This type of sentence may or may not have citations. Citations are likely to be in parenthesis and not part of the syntax of the sentence. Examples are provided below in Figure 1.

1. *Most of the previous works conduct structure alignment with complex , hierarchical structures , such as phrase structures (e.g. , Kaji , Kida & Morimoto , 1992) , or dependency structures (e.g. , Matsumoto et al. 1993 ; Grishman , 1994 ; Meyers , Yanharber & Grishman 1996 ; Watanabe , Kurohashi & Aramaki 2000).*
2. *Then the correspondent structures in different languages are aligned (e.g. , Kaji , Kida & Morimoto 1992 ; Matsumoto et al. 1993 ; Grishman 1994 ; Meyers , Yanharber & Grishman 1996 ; Watanabe , Kurohashi & Aramaki 2000).*
3. *In general , current approaches to NE identification (e.g. Chen , 1997) usually contain two separate steps : word segmentation and NE identification.*
4. *These models are trained on a parallel corpus of long source sentences and their target compressions.*
5. *Syntactic structure matching has been applied to passage retrieval (Cui et al. , 2005) and answer extraction (Shen and Klakow . 2006) .*

Figure 1 Background Example Sentences

Other Work (cited work)

This type of sentence talks specifically about a cited work and the citation is usually part of the syntax of the sentence or follows on from a sentence that contained such a citation.

Author's Work

This type of sentence talks about the author's work in this paper only.

Multi-type Subjects

There will be sentences that have multiple subjects such as those that compare cited works or compare the author's work to a cited work or say how the author's work differs from the field in general.

Labelling the Sentence

After deciding on the subject use the steps below to help decide on an annotation label that best describes the intention of the sentence, based on the available annotation labels (Figure 2).

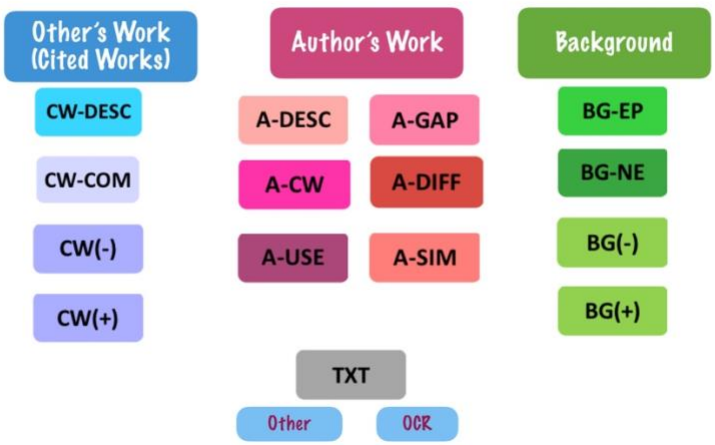


Figure 2 Annotation Labels

Sentence Subject – Background

If you decide on a sentence subject of '*Background*'. You have the following choices:

Is the sentence a description of state in the field, describing/listing generally a known method or general knowledge?

- Does is make reference to a shortcoming or gap in the background/field? (see example Figure 3)
  - YES -> then label **BG(-)**

1. *Finally , the machine learning-based model has also been investigated . current models of this type are based on supervised approaches [Ittycheriah et al. 2001 ; Ng et al. 2001 ; Suzuki et al. 2002 ; and Sasaki et al. 2005] that are heavily dependent on hand-tagged question-answer training pairs , which are not readily available.*

Figure 3 Background Shortcoming/Gap Example

- Does is make reference to a positive in the background/field? (see example Figure 4)
- YES -> then label **BG(+)**

1. *Recently , statistical NERs have achieved results that are comparable to hand-coded systems.*

Figure 4 Background Positive Example

- It is a background sentence that provides a citation?
  - Yes -> then label **BG-EP**
  - No - > then label **BG-NE**

Sentence Subject – Other's Work (Cited Work)

If you decide on a sentence subject of 'Other's Work'. You have the following choices:

- The author discusses a shortcoming, limitation or gap that the cited work does not address.
    - YES -> then label **CW(-)**
- The author discusses a positive about the cited work .
    - YES -> then label **CW(+)**
- The author compares two cited works.
    - YES -> then label **CW-COMP**
- The sentence is a description/detail of the work only.
    - YES -> then label **CW-DESC**

Sentence Subject – Author's Work

If you decide on a sentence subject of 'Author's Work'. You have the following choices:

Firstly, the author may simply say that their work is different but give no details or they may say to the best of their knowledge they are the only people who have done this but again no detail. (See Figure 5 for example sentences)

- YES -> then label **A-DIFF**

1. *Our work differs from previous approaches in two key respects.*
2. *However , we differ in two important respect .*
3. *There are some algorithmic differences between these papers and ours.*
4. *To the best of our knowledge this is the first paper providing a probabilistic generative , history-based generation model.*

Figure 5 A-Diff Example Sentences

An author may specifically discuss or mention that they address a gap:- there is something new, novel or better about their work. They are not directly comparing it to something, just saying they do something new or better. You should only mark a sentence as A-GAP if it is clear from the text that it is novel or a contribution. (Note if they are ***comparing*** to the field or an other's work – read on to A-CW labels) (See Figure 6 for example sentences.)

- YES -> then label **A-GAP**

1. *However , because our method does not require sentence alignments , it can be applied for wider applicable domains.*
2. *However , since our method caught extracting the translation pairs as the approach of the statistical machine learning , it could be expected to improve the performance by adding new features to the translation model.*
3. *In addition , if learning the translation model for the training samples is done once with our method , the model need not be learned again for new samples although it needs the positive and negative samples.*

Figure 6 A-GAP example sentences

When authors talk about their work being different the author may directly compare their work to someone else's or the field in general. The next three labels account for this. Sometimes a sentence may be a combination, e.g. they may say they are similar or use another work but then in the same sentence they say the differ, in this case choose the A-CW (for author's work compared to other work). They could also say something about a limitation of a citation - CW(-) then say how their work differs in a sentence, again label this as A-CW. Choose a label that captures the author's work is compared rather than the single CW type label.

If the author compares their work saying it differs to what has been done in general or to a specific cited works(s) (See examples Figure 7)

- YES -> then label **A-CW**

1. *In contrast to Kaisser (2006) , we model the semantic role assignment and answer extraction tasks numerically , thereby alleviating the coverage problems encountered previously .*
2. *We differ from the supervised techniques described , in which a large number of hand-tagged training pairs are shared by all of the test questions .*
3. *First , our generation algorithm is more powerful , performing complex tree transformations , whereas McDonald (1990) only considers simple word deletion .*
4. *Different from his work , foreign syntactic knowledge is introduced into the synchronous grammar rules in our method to restrict the arbitrary phrase reordering .*
5. *Though their model is also based on hierarchical Dirichlet processes and is similar to ours , they present a different inference algorithm which is based on sampling .*

Figure 7 A-CW Sentence Examples

A sentence that compares saying that the author build/use/are inspired by the work or works in general (see examples Figure 8)

- YES -> then label **A-USE**

1. *This method is also adopted in our system for nonpeer phrase reordering .*
2. *This approach is inspired by methods in the topic modeling literature , such as Latent Dirichlet Allocation (LDA) (Blei et al. , 2003) , where topics are treated as hidden variables that govern the distribution of words in a text .*

Figure 8 A-Use sentence examples

A sentence compares saying that their own work (i.e. the work in this paper) is similar to a cited work or works to the field in general. Sometimes an author notes the similarity at the start of the sentence but ultimately the author says they differ – label this as **A-CW** rather than this label(see example sentences Figure 9)

- YES -> then label **A-SIM**

1. *Like our method , researches which are not based on the assumption of the sentence alignments for parallel corpora have been done (Kaji and Aizono , 1996 ; Tanaka and Iwasaki , 1996 ; Fung , 1997) .*
2. *In line with previous work , our method exploits syntactic information in the form of dependency relation paths together with FrameNet-like semantic roles to smooth lexical and syntactic divergences between question and answer sentences .*

Figure 9 A-Sim Example Sentences

Finally, an author sometimes just gives a description of what they do and there is no indication if this is different, something new/better or that it compares to something done before.

- YES -> then label **A-DESC**

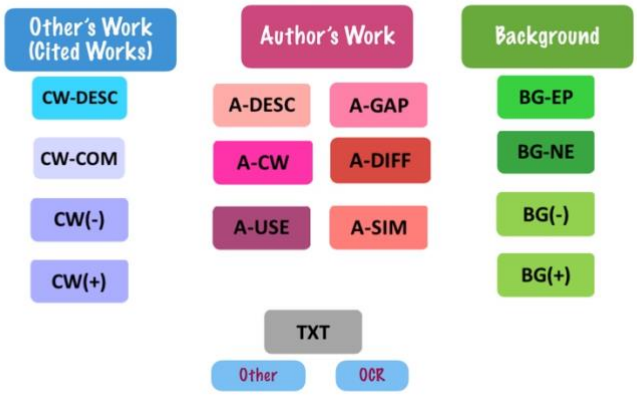
1. *In this framework , we integrate Chinese word segmentation and NE identification into a unified framework using a class-based language model (LM) .*
2. *We only combine top-one hypothesis from each system , and did not apply system confidence measure and minimum error rate training to tune system combination weights .*
3. *Hybrid indexing allows us to compute semantic cohesion score rather than the lexical cohesion score based on word repetitions .*

Figure 10 A-DESC Example Sentences

Other Label Categories

We have several categories to deal sentences that do not fit into any of the above.

- **OCR** – if there has been a problem with rendering the text from the PDF such that a label cannot be chosen, select the label OCR, if it is a minor OCR and you can choose a label please do so
- **TXT**- these are sentences where the author says *in section 6 we discuss X* or something similar. It is a pointer to somewhere in the text. There are also instances of things like *this can be found in table 1*.
- **Other** – Does not belong in any other category



Bibliography

Abdalla, R. M. and Teufel, S. (2006). A bootstrapping approach to unsupervised detection of cue phrase variants. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 921–928. Association for Computational Linguistics.

Abel, S., Kitto, K., Knight, S., and Buckingham Shum, S. (2018). Designing personalised, automated feedback to develop students' research writing skills. In *Open Oceans: Learning without borders. Proceedings ASCILITE 2018*, pages 15–24.

Afantenos, S., Denis, P., Muller, P., and Danlos, L. (2010). Learning recursive segments for discourse parsing. *arXiv preprint arXiv:1003.5372*.

Aitchison, C., Catterall, J., Ross, P., and Burgin, S. (2012). 'tough love and tears': learning doctoral writing in the sciences. *Higher Education Research & Development*, 31(4):435–447.

Aldayel, A. and Magdy, W. (2019). Your stance is exposed! analysing possible factors for stance detection on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–20.

Angrosh, M. A., Cranefield, S., and Stanger, N. (2012). Context identification of sentences in research articles: Towards developing intelligent tools for the research community. *Natural Language Engineering*, 19(04):481–515.

Anthony, L. and V. Lashkia, G. (2003). Mover: A Machine Learning Tool to Assist in the Reading and Writing of Technical Papers. *Professional Communication, IEEE Transactions on*, 46:185–193.

Artstein, R. and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Asadi, N., Badie, K., and Mahmoudi, M. T. (2019). Automatic zone identification in scientific papers via fusion techniques. *Scientometrics*, 119(2):845–862.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*, pages 81–87, Portland, OR, USA. Association for Computational Linguistics.

Athar, A. (2014). Sentiment analysis of scientific citations. Technical Report UCAM-CL-TR-856, University of Cambridge, Computer Laboratory.

Barker, E. and Gaizauskas, R. (2016). Summarizing multi-party argumentative conversations in reader comment on news. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 12–20.

Batista-Navarro, R. T. and Ananiadou, S. (2011). Building a coreference-annotated corpus from the domain of biochemistry. In *Proceedings of BioNLP 2011 Workshop*, pages 83–91, Portland, Oregon, USA. Association for Computational Linguistics.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*, volume 23. John Benjamins Publishing.

Biber, D., Conrad, S., and Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3):371–405.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., Powley, B., Radev, D., and Tan, Y. F. (2008). The ACL Anthology Reference Corpus. In *LREC*, pages 1–6.

Boote, D. and Beile, P. (2016). Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. *Educational Researcher*, 34(6):3–15.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bridges, D. (2009). Research quality assessment in education: impossible science, possible art? *British Educational Research Journal*.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25:27–36.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*.

Cambre, J., Klemmer, S., and Kulkarni, C. (2018). Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 294. ACM.

Casey, A., Webber, B., and Głowacka, D. (2019a). Can models of author intention support quality assessment of content? In *BIRNDL 2019*.

Casey, A. J., Webber, B., and Głowacka, D. (2019b). A Framework for Annotating 'Related Works', to Support Feedback to Novice Writers. In *LAW '13: Proceedings of the Linguistic Annotation Workshop*. Association for Computational Linguistics.

Casey, A. J., Webber, B., and Głowacka, D. (2019c). Classifying Author Intention for Writer Feedback in Related Works. In *RANLP: Recent Advances in Natural Language Processing 2019*. Association for Computational Linguistics.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, 23(4):397–423.

Cotos, E. (2009). Designing an intelligent discourse evaluation tool: Theoretical, empirical, and technological considerations. *Developing and Evaluating Language Learning Materials*, page 103–127.

Cotos, E. (2014). *Genre-based automated writing evaluation for L2 research writing: From design to evaluation and enhancement*. Springer.

Cotos, E. and Pendar, N. (2016). Discourse classification into rhetorical functions for awe feedback. *calico journal*, 33(1):92–116.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press, Princeton.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dgani, Y., Greenspan, H., and Goldberger, J. (2018). Training a neural network based on unreliable human annotation of medical images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 39–42. IEEE.

Egan, C., Siddharthan, A., and Wyner, A. (2016). Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143.

Feltrim, V. D., Teufel, S., das Nunes, M. G. V., and Aluísio, S. M. (2006). Argumentative zoning applied to critiquing novicesâ^{TM} scientific abstracts. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 233–246. Springer.

Fisas, B., Ronzano, F., and Saggion, H. (2016). A multi-layered annotated corpus of scientific papers. In *LREC 2016*.

Fisas, B., Saggion, H., and Ronzano, F. (2015). On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th linguistic annotation workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gasperin, C. V. (2009). Statistical anaphora resolution in biomedical texts. Technical Report UCAM-CL-TR-764, University of Cambridge, Computer Laboratory.

Ghosh, D., Khanam, A., Han, Y., and Muresan, S. (2016). Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Gibson, A., Aitken, A., Sándor, A., Buckingham Shum, S., Tsingos-Lucas, C., and Knight, S. (2017). Reflective writing analytics for actionable feedback. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 153–162, New York, NY, USA. ACM.

Gkatzia, D., Lemon, O., and Rieser, V. (2016). Natural language generation enhances human decision-making with uncertain information. *arXiv preprint arXiv:1606.03254*.

Glanzel, W., Debackere, K., Thijs, B., and Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*.

Gogolin, I. and Stumm, V. (2014). The eerqi peer review questionnaire–from the development of 'intrinsic indicators' to a tested instrument. In *Assessing Quality in European Educational Research*, pages 107–120. Springer.

Gorinski, P. J., Wu, H., Grover, C., Tobin, R., Talbot, C., Whalley, H., Sudlow, C., Whiteley, W., and Alex, B. (2019). Named entity recognition for electronic

health records: a comparison of rule-based and machine learning approaches. *arXiv preprint arXiv:1903.03985*.

Green, N. (2017). Manual Identification of Arguments with Implicit Conclusions Using Semantic Rules for Argument Mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 73–78. Association for Computational Linguistics.

Grosz, B. J., Joshi, A. K., and Weinstein, S. (1995). Centering: A Framework for Modelling the Local Coherence of Discourse,. Technical report, US Dept of the Army, Funding, Fort Belvoir, VA.

Gwet, K. (2014). *Handbook of Inter-Rater Reliability*. Advanced Analytics, LLC.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2017). The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. *arXiv preprint arXiv:1708.01425*.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). Semeval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772.

Harmon, J. E. and Gross, A. G. (2010). *The Craft of Scientific Communication*. The University of Chicago Press.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Heffernan, K. and Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2):1367–1382.

Hockly, N. (2019). *Automated writing evaluation*.

Hussein, M. A., Hassan, H., and Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, 5:e208.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Hyland, K. (2015). Genre, discipline and identity. *Journal of English for Academic Purposes*, 19(C):32–43.

Isard, A. and Knox, J. (2016). Automatic generation of student report cards. In *Proceedings of the 9th International Natural Language Generation conference*, pages 207–211.

Jackson, P. and Moulinier, I. (2002). Natural language processing for online applications: Text retrieval. *Extraction and Categorization*.

Jespersen, O. (1924). *The philosophy of grammar*. London: Allen and Unwi.

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., Moore, A. L., and Biswas, G. (2018). Expert feature-engineering vs. deep neural networks: which is better for sensor-free affect detection? In *International Conference on Artificial Intelligence in Education*, pages 198–211. Springer.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., and Jurafsky, D. (2018). Measuring the Evolution of a Scientific Field through Citation Frames. *Transactions of the Association of Computational Linguistics*, 6:391–406.

Kamler, B. and Thomson, P. (2006). *Helping doctoral students write: Pedagogies for supervision*. Routledge.

Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and trends in Information Retrieval*, 3(1—2):1–224.

Kim, Y. and Webber, B. (2006). Automatic reference resolution in astronomy articles. In *Proc. of 20th International CODATA Conference, Beijing, China*.

Kincaid, J. P., Fishburne, J., Robert P, R., Richard L, C., and S, B. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report, US Dept of the Navy, Funding, Fort Belvoir, VA.

Kirschner, C., Eckle-Kohler, J., and Gurevych, I. (2015). Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 1–11.

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., and Wang, X. (2018). Designing Academic Writing Analytics for Civil Law Student Self-Assessment. *International Journal of Artificial Intelligence in Education*, 28(1):1–28.

Krosnick, J. and Presser, S. (2010). Question and questionnaire design. marsden pv and wright jd (eds.) handbook of survey research, vol. 2.

Kwitt, R., Hegenbart, S., Rasiwasia, N., Vécsei, A., and Uhl, A. (2014). Do we need annotation experts? a case study in celiac disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 454–461. Springer.

Lamprier, S., Amghar, T., Levrat, B., and Saubion, F. (2007). On evaluation methodologies for text segmentation algorithms. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 19–26. IEEE.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lauscher, A., Glavaš, G., and Eckert, K. (2018). Arguminsci: A tool for analyzing argumentation and rhetorical aspects in scientific writing. Association for Computational Linguistics.

Lawrence, N. and Cortes, C. (2014). The NIPS experiment.

Liakata, M., Saha, S., Dobnik, S., and Batchelor, C. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Lin, Z., Ng, H. T., and Kan, M.-Y. (2014). A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184.

Litman, D. J. (1996). Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

Liu, J., Xu, Y., and Zhao, L. (2019a). Automated essay scoring based on two-stage learning. *arXiv preprint arXiv:1901.07744*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Louis, A. and Nenkova, A. (2012). A coherence model based on syntactic patterns. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, page 1157–1168, USA. Association for Computational Linguistics.

Lucas, C., Gibson, A., and Buckingham Shum, S. (2018). Utilization of a novel online reflective learning tool for immediate formative feedback to assist pharmacy students reflective writing skills. *Am J Pharm Educ*, 83(6).

MacRoberts, M. H. and MacRoberts, B. R. (1984). The negational reference: or the art of dissembling. *Social Studies of Science*, 14(1):91–94.

Maher, M. A., Feldon, D. F., Timmerman, B. E., and Chao, J. (2014). Faculty perceptions of common challenges encountered by novice doctoral writers. *Higher Education Research & Development*, 33(4):699–711.

Mangiafico, S. (2019). How should we analyze likert item data? *Journal of National Association of County Agricultural Agents (NACAA)*, 12(2).

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Maxwell, J. A. (2006). Literature Reviews of, and for, Educational Research: A Commentary on Boote and Beile's "Scholars Before Researchers". *Educational Researcher*, 35(9):28–31.

Merity, S., Murphy, T., and Curran, J. R. (2009). Accurate argumentative zoning with maximum entropy models. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 19–26.

Mikolov, T., Chen, K., Corrado, G., Dean, J., Sutskever, L., and Zweig, G. (2013). word2vec. *URL https://code. google. com/p/word2vec*.

Miltsakaki, E. and Kukich, K. (2004). Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Mizuta, Y. and Collier, N. (2004). Zone identification in biology articles as a basis for information extraction. In *COLING 2004 International Joint workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP) 2004*, pages 29–35, Geneva, Switzerland. COLING.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Moilanen, K.-H. and Pulman, S. G. (2016). System and method for analysing natural language. US Patent 9,336,205.

Müller, C. and Strube, M. (2006). Multi-level annotation of linguistic data with MMAX2. In Braun, S., Kohn, K., and Mukherjee, J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214. Peter Lang, Frankfurt a.M., Germany.

Nadeau, C. and Bengio, Y. (1999). Inference for the generalization error. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 307–313, Cambridge, MA, USA. MIT Press.

Nadeem, F., Nguyen, H., Liu, Y., and Ostendorf, M. (2019). Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493.

Nash, P. and Shaffer, D. W. (2011). Mentor modeling: The internalization of modeled professional thinking in an epistemic game. *Journal of Computer Assisted Learning*, 27(2):173–189.

Nguyen, H. V. and Litman, D. J. (2018). Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, memory, and cognition*, 14(3):510.

Ong, N., Litman, D., and Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28.

Oppenheim, C. and Renn, S. P. (1978). Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):225–231.

O'Rourke, S. T. and Calvo, R. A. (2009). Visualizing paragraph closeness for academic writing support. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*, pages 688–692.

Paltridge, B. and Starfield, S. (2007). *Thesis and Dissertation Writing in a Second Language*. Routledge.

Paré, A. (2010). Making sense of supervision: Deciphering feedback. *The Routledge doctoral student's companion: Getting to grips with research in education and the social sciences*, pages 107–115.

Passonneau, R. J. and Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications, inc.

Peldszus, A. and Stede, M. (2015). Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *CoRR*, abs/1802.05365.

Pevzner, L. and Hearst, M. A. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Pitler, E. and Nenkova, A. (2008). Revisiting readability. In *the Conference*, pages 186–195, Morristown, NJ, USA. Association for Computational Linguistics.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ritchie, A., Teufel, S., and Robertson, S. (2006). How to find better index terms through citations. In *Proceedings of the workshop on how can computational linguistics improve information retrieval?*, pages 25–32. Association for Computational Linguistics.

Ritchie, A., Teufel, S., and Robertson, S. (2008). Using terms from citations for IR: some first results. In *European Conference on Information Retrieval*, pages 211–221. Springer.

Rösiger, I. and Teufel, S. (2014). Resolving coreferent and associative noun phrases in scientific text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 45–55.

Ross, P., Burgin, S., Aitchison, C., and Catterall, J. (2011). Research writing in the sciences: Liminal territory and high emotion. *Journal of Learning Design*, 4.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, 18(2):119–144.

Sándor, Á., Kaplan, A., and Rondeau, G. (2006). Discourse and citation analysis with concept-matching. In *International Symposium: Discourse and document (ISDD)*, pages 15–16.

Sándor, Á. and Vorndran, A. (2014). *Highlighting Salient Sentences for Reading Assistance*, pages 43–55. Springer Fachmedien Wiesbaden, Wiesbaden.

Schäfer, U., Spurk, C., and Steffen, J. (2012). A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology. In *Proceedings of COLING 2012: Posters*, pages 1059–1070, Mumbai, India. The COLING 2012 Organizing Committee.

Schwarm, S. E. and Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In *the 43rd Annual Meeting*, pages 523–530, Morristown, NJ, USA. Association for Computational Linguistics.

Shaffer, D. W., Collier, W., and Ruis, A. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, 3(3):9–45.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., and Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media*, 1(2).

Shum, S. B., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., and McWilliams, M. (2016). Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 213–222.

Si, L. and Callan, J. (2001). A statistical model for scientific readability. In *the tenth international conference*, pages 574–576, New York, New York, USA. ACM Press.

Siddharthan, A. and Teufel, S. (2007). Whose idea was this, and why does it matter? attributing scientific work to citations. In *Human language technologies 2007: The conference of the North American chapter of the Association for Computational Linguistics; proceedings of the main conference*, pages 316–323.

Siddiqua, U. A., Chy, A. N., and Aono, M. (2018). Stance detection on microblog focusing on syntactic tree representation. In *International Conference on Data Mining and Big Data*, pages 478–490. Springer.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Song, Y., Heilman, M., Beigman Klebanov, B., and Deane, P. (2014). Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

Sporleder, C. and Lapata, M. (2005). Discourse chunking and its application to sentence compression. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 257–264. Association for Computational Linguistics.

Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Sumner, M., Frank, E., and Hall, M. A. (2005). Speeding up logistic model tree induction. *PKDD*, LNCS 3721:675–683.

Swales, J. (1981). *Aspects of article introductions*. Language Studies Unit. University of Aston in Birmingham.

Swales, J. (1990). *Genre Analysis: English in academic and research settings*. Cambridge University Press.

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. PhD thesis, University of Edinburgh.

Teufel, S. and Kan, M.-Y. (2009). Robust argumentative zoning for sensemaking in scholarly documents. In *Advanced language technologies for digital libraries*, pages 154–170. Springer.

Teufel, S. and Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445.

Teufel, S., Siddharthan, A., and Batchelor, C. (2009). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502, Singapore. Association for Computational Linguistics.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87, Sydney, Australia. Association for Computational Linguistics.

Teufel, S., Siddharthan, A., and Tidhar, D. (2006b). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Thompson, P. and Tribble, C. (2001). Looking at Citations: Using Corpora in English for Academic Purposes. *Language Learning Technology*, 5(3):91 – 105.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018a). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2018b). The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press.

Vargha, A. and Delaney, H. D. (2000). A critique and improvement of the cl common language effect size statistics of mcgraw and wong. *Journal of Educational and Behavioral Statistics*, 25(2):101–132.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Versley, Y. (2008). Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, 6(3-4):333–353.

Walker, R. and da Silva, P. R. (2014). Emerging trends in peer review-a survey. *Frontiers in Neuroscience*.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

Weinstock, M. (1971). Citation Indexes. Encyclopedia of Library and Information Science. volume 5. eds. a. kent & h. lancour.

Yang, Y., Agarwal, O., Tar, C., Wallace, B. C., and Nenkova, A. (2019). Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. *arXiv preprint arXiv:1905.07791*.